Horizon 2020

Call: H2020-FETHPC-2016-2017
(FET Proactive – High Performance Computing)

Topic: FETHPC-02-2017

Type of action: RIA
(Research and Innovation action)

Proposal number: 800929

Proposal acronym: DIO

Deadline Id: H2020-FETHPC-2017

Table of contents

How to fill in the forms

The administrative forms must be filled in for each proposal using the templates available in the submission system. Some data fields in the administrative forms are pre-filled based on the previous steps in the submission wizard.

1 - General information

Topic	FETHPC-02-2017
Call Identifier	H2020-FETHPC-2016-2017
Type of Action	RIA
Deadline Id	H2020-FETHPC-2017

Acronym	DIO		
Proposal title*	High Density I/O Path for Fast Storage Devices and Multi-tier Hierarchies *Note that for technical reasons, the following characters are not accepted in the Proposal Title and will be removed: < > " &*		
Duration in months	*36*		
Fixed keyword 1	*High performance computing*	Add	
Fixed keyword 2	*Computer systems, parallel/distributed systems, sensor network*	Add	Remove
Fixed keyword 3	*System Software*	Add	Remove
Fixed keyword 4	*Scalability*	Add	Remove
Free keywords	*Enter any words you think give extra detail of the scope of your proposal (max 200 characters with spaces).*		

Proposal ID **800929** *Acronym* **DIO**

Abstract

DIO will design, prototype, and demonstrate an I/O software stack for deep storage hierarchies that incorporates the following technology innovations and methodology:
(a) Global I/O address space via partitioning, placing, and indexing data on devices, based on multi-level write-optimized data structures that have emerged in the area of analytics and have the ability to inherently adapt to different device technologies at each level/tier.
(b) Novel memory-mapped I/O approach for fast devices, such as NVMe and NVM, to organize and access data directly on devices, which will eliminate all software path overhead in the common I/O path, both for caching (hits) and data access (kernel crossings). DIO will use transparent data replication at the mmap level to replicate data over fast remote devices using RDMA and failure atomicity based on copy-on-write.
(c) Use of FPGA acceleration for near-device processing both for reducing I/O overheads, such as replication coding, but also for offloading application-induced data transformations that can be performed during data replacement.
(d) Support legacy APIs for existing applications but will also explore new capabilities, especially with relation to device (tier) technology and in-I/O-path acceleration.
(e) Use of a key-value based abstraction for locating and accessing data, which will allow sharing, access, and prioritization at fine grain and on a per-pair basis.
(f) Mechanisms and policies for allocating resources to concurrent applications and diverse workloads that exhibit high-degrees of parallelism and require dynamic data placement across tiers to achieve both deterministic application QoS and high system utilization.
(g) Converged rack-scale architecture for bringing the proposed technology to next-generation systems.

Remaining characters 202

Has this proposal (or a very similar one) been submitted in the past 2 years in response to a call for proposals under Horizon 2020 or any other EU programme(s)? ○ Yes ● No

Proposal ID **800929** *Acronym* **DIO**

Declarations

1) The coordinator declares to have the explicit consent of all applicants on their participation and on the content of this proposal.	☒
2) The information contained in this proposal is correct and complete.	☒
3) This proposal complies with ethical principles (including the highest standards of research integrity — as set out, for instance, in the European Code of Conduct for Research Integrity — and including, in particular, avoiding fabrication, falsification, plagiarism or other research misconduct).	☒

4) The coordinator confirms:

- to have carried out the self-check of the financial capacity of the organisation on http://ec.europa.eu/research/participants/portal/desktop/en/organisations/lfv.html or to be covered by a financial viability check in an EU project for the last closed financial year. Where the result was "weak" or "insufficient", the coordinator confirms being aware of the measures that may be imposed in accordance with the H2020 Grants Manual (Chapter on Financial capacity check); or	○
- is exempt from the financial capacity check being a public body including international organisations, higher or secondary education establishment or a legal entity, whose viability is guaranteed by a Member State or associated country, as defined in the H2020 Grants Manual (Chapter on Financial capacity check); or	●
- as sole participant in the proposal is exempt from the financial capacity check.	○

5) The coordinator hereby declares that each applicant has confirmed:

- they are fully eligible in accordance with the criteria set out in the specific call for proposals; and	☒
- they have the financial and operational capacity to carry out the proposed action.	☒

The coordinator is only responsible for the correctness of the information relating to his/her own organisation. Each applicant remains responsible for the correctness of the information related to him/her and declared above. Where the proposal to be retained for EU funding, the coordinator and each beneficiary applicant will be required to present a formal declaration in this respect.

According to Article 131 of the Financial Regulation of 25 October 2012 on the financial rules applicable to the general budget of the Union (Official Journal L 298 of 26.10.2012, p. 1) and Article 145 of its Rules of Application (Official Journal L 362, 31.12.2012, p.1) applicants found guilty of misrepresentation may be subject to administrative and financial penalties under certain conditions.

Personal data protection

The assessment of your grant application will involve the collection and processing of personal data (such as your name, address and CV), which will be performed pursuant to Regulation (EC) No 45/2001 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data. Unless indicated otherwise, your replies to the questions in this form and any personal data requested are required to assess your grant application in accordance with the specifications of the call for proposals and will be processed solely for that purpose. Details concerning the purposes and means of the processing of your personal data as well as information on how to exercise your rights are available in the privacy statement. Applicants may lodge a complaint about the processing of their personal data with the European Data Protection Supervisor at any time.

Your personal data may be registered in the Early Detection and Exclusion system of the European Commission (EDES), the new system established by the Commission to reinforce the protection of the Union's financial interests and to ensure sound financial management, in accordance with the provisions of articles 105a and 108 of the revised EU Financial Regulation (FR) (Regulation (EU, EURATOM) 2015/1929 of the European Parliament and of the Council of 28 October 2015 amending Regulation (EU, EURATOM) No 966/2012) and articles 143 - 144 of the corresponding Rules of Application (RAP) (COMMISSION DELEGATED REGULATION (EU) 2015/2462 of 30 October 2015 amending Delegated Regulation (EU) No 1268/2012) for more information see the Privacy statement for the EDES Database).

List of participants

#	Participant Legal Name	Country
1	FOUNDATION FOR RESEARCH AND TECHNOLOGY HELLAS	Greece
2	BARCELONA SUPERCOMPUTING CENTER - CENTRO NACIONAL DE SUPERCOMPUTACION	Spain
3	SCIENCE AND TECHNOLOGY FACILITIES COUNCIL	United Kingdom
4	BULL SAS	France
5	JOHANNES GUTENBERG-UNIVERSITAT MAINZ	Germany
6	INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS	Greece
7	CYBELETECH SAS	France

2 - Administrative data of participating organisations

PIC	***Legal name***
999995893	*FOUNDATION FOR RESEARCH AND TECHNOLOGY HELLAS*

Short name: FORTH

Address of the organisation

Street N PLASTIRA STR 100

Town HERAKLION

Postcode 70013

Country Greece

Webpage www.forth.gr

Legal Status of your organisation

Research and Innovation legal statuses

Public body ....no

Legal person .... yes

Non-profit ....yes

International organisation ....no

International organisation of European interest ......no

Secondary or Higher education establishment ....no

Research organisation ....yes

Enterprise Data

SME self-declared status..................................11/05/2016 - no

SME self-assessment .... unknown

SME validation sme... 25/09/2008 - no

Based on the above details of the Beneficiary Registry the organisation is not an SME (small- and medium-sized enterprise) for the call.

Department(s) carrying out the proposed work

Department 1

Department name	Institute of Computer Science / CARV laboratory	□ not applicable
	☒ Same as organisation address	
Street	N PLASTIRA STR 100	
Town	HERAKLION	
Postcode	70013	
Country	Greece	

Dependencies with other proposal participants

Character of dependence	***Participant***	

Person in charge of the proposal

The name and e-mail of contact persons are read-only in the administrative form, only additional details can be edited here. To give access rights and basic contact details of contact persons, please go back to Step 4 of the submission wizard and save the changes.

Title	Prof.	Sex	● Male ○ Female
First name	**Angelos**	Last name	**BILAS**
E-Mail	**bilas@ics.forth.gr**		
Position in org.	Associated Faculty Member		
Department	Institute of Computer Science / CARV Laboratory		□ Same as organisation
	☒ Same as organisation address		
Street	N PLASTIRA STR 100		
Town	HERAKLION	Post code	70013
Country	Greece		
Website	http://www.ics.forth.gr/~bilas		
Phone 1	+30 2810391669	Phone 2	+xxx xxxxxxxxx
Fax	+30 2810391661		

Other contact persons

First Name	***Last Name***	***E-mail***	***Phone***
Christos	Kozanitis	kozanitis@ics.forth.gr	
Manolis	Marazakis	maraz@ics.forth.gr	
Ioannis	Stratakis	gstra@ics.forth.gr	

PIC	***Legal name***
999655520	*BARCELONA SUPERCOMPUTING CENTER - CENTRO NACIONAL DE SUPERCOMPUTACION*

Short name: BSC

Address of the organisation

Street	Calle Jordi Girona 31
Town	BARCELONA
Postcode	08034
Country	Spain
Webpage	www.bsc.es

Legal Status of your organisation

Research and Innovation legal statuses

Public body ....yes

Legal person .... yes

Non-profit ....yes

International organisation ....no

International organisation of European interest ......no

Secondary or Higher education establishment ....no

Research organisation ....yes

Enterprise Data

SME self-declared status..................................01/03/2005 - no

SME self-assessment .... unknown

SME validation sme.. unknown

Based on the above details of the Beneficiary Registry the organisation is not an SME (small- and medium-sized enterprise) for the call.

Department(s) carrying out the proposed work

Department 1

Department name	Computer Science	□ not applicable
	☒ Same as organisation address	
Street	Calle Jordi Girona 31	
Town	BARCELONA	
Postcode	08034	
Country	Spain	

Dependencies with other proposal participants

Character of dependence	***Participant***	

Person in charge of the proposal

The name and e-mail of contact persons are read-only in the administrative form, only additional details can be edited here. To give access rights and basic contact details of contact persons, please go back to Step 4 of the submission wizard and save the changes.

Title	Prof.	Sex	● Male ○ Female
First name	**Toni**	Last name	**Cortes**
E-Mail	**toni.cortes@bsc.es**		
Position in org.	Group Leader		
Department	Computer Science		□ Same as organisation
	☒ Same as organisation address		
Street	Calle Jordi Girona 31		
Town	BARCELONA	Post code	08034
Country	Spain		
Website	www.bsc.es		

Phone 1	+34 934137966	Phone 2	+34 934137569
Fax	+xxx xxxxxxxxx		

Other contact persons

First Name	***Last Name***	***E-mail***	***Phone***
Ramón	Nou	ramon.nou@bsc.es	+34 934137569
Isabel	Martinez	isabel.martinez@bsc.es	+34 934137075
Alberto	Miranda	alberto.miranda@bsc.es	

PIC	***Legal name***
999980179	*SCIENCE AND TECHNOLOGY FACILITIES COUNCIL*

Short name: STFC

Address of the organisation

Street Polaris House North Star Avenue

Town SWINDON

Postcode SN2 1SZ

Country United Kingdom

Webpage www.scitech.ac.uk

Legal Status of your organisation

Research and Innovation legal statuses

Public body ....yes Legal person .... yes

Non-profit ....yes

International organisation ....unknown

International organisation of European interest ......unknown

Secondary or Higher education establishment ....unknown

Research organisation ....yes

Enterprise Data

SME self-declared status..................................01/04/2007 - no

SME self-assessment .... unknown

SME validation sme.. unknown

Based on the above details of the Beneficiary Registry the organisation is not an SME (small- and medium-sized enterprise) for the call.

Department(s) carrying out the proposed work

Department 1

Department name	Hartree Centre	□ not applicable
	□ Same as organisation address	
Street	Sci-Tech Daresbury	
Town	Warrington	
Postcode	WA4 4AD	
Country	United Kingdom	

Dependencies with other proposal participants

Character of dependence	***Participant***	

Person in charge of the proposal

The name and e-mail of contact persons are read-only in the administrative form, only additional details can be edited here. To give access rights and basic contact details of contact persons, please go back to Step 4 of the submission wizard and save the changes.

Title	Dr.	Sex	● Male ○ Female
First name	**Milos**	Last name	**Puzovic**
E-Mail	**milos.puzovic@stfc.ac.uk**		
Position in org.	Research Scientist		
Department	Future Technologies		□ Same as organisation
	□ Same as organisation address		
Street	Sci-Tech Daresbury		
Town	Warrington	Post code	WA4 4AD
Country	United Kingdom		
Website	http://www.stfc.ac.uk/hartree		

Phone 1	+44 1925 603 396	Phone 2	+xxx xxxxxxxxx	Fax	+xxx xxxxxxxxx

Other contact persons

First Name	***Last Name***	***E-mail***	***Phone***
Michael	Bane	michael.bane@stfc.ac.uk	
Charles	Moulinec	charles.moulinec@stfc.ac.uk	

PIC	***Legal name***
996058081	*BULL SAS*

Short name: BULL

Address of the organisation

Street	RUE JEAN JAURES 68
Town	LES CLAYES SOUS BOIS
Postcode	78340
Country	France
Webpage	www.bull.com

Legal Status of your organisation

Research and Innovation legal statuses

Public body ....no

Legal person .... yes

Non-profit ....no

International organisation ....no

International organisation of European interest ......no

Secondary or Higher education establishment ....no

Research organisation ....no

Enterprise Data

SME self-declared status................................. unknown

SME self-assessment .... unknown

SME validation sme...................................... unknown

Based on the above details of the Beneficiary Registry the organisation is not an SME (small- and medium-sized enterprise) for the call.

Department(s) carrying out the proposed work

Department 1

Department name: HPC and Server Design Department

□ not applicable

☒ Same as organisation address

Street: RUE JEAN JAURES 68

Town: LES CLAYES SOUS BOIS

Postcode: 78340

Country: France

Dependencies with other proposal participants

Character of dependence	***Participant***	

Person in charge of the proposal

The name and e-mail of contact persons are read-only in the administrative form, only additional details can be edited here. To give access rights and basic contact details of contact persons, please go back to Step 4 of the submission wizard and save the changes.

Title	Dr.	Sex	● Male ○ Female
First name	**Huy Nam**	Last name	**Nguyen**
E-Mail	**huy-nam.nguyen@bull.net**		
Position in org.	Head of MVS		
Department	HPC and Server Design Department		□ Same as organisation
	☒ Same as organisation address		
Street	RUE JEAN JAURES 68		
Town	LES CLAYES SOUS BOIS	Post code	78340
Country	France		
Website	www.bull.com		

Phone 1	+33130806019	Phone 2	+xxx xxxxxxxxx	Fax	+33130806157

Proposal ID	**800929**	*Acronym*	**DIO**	*Short name*	**JGU MAINZ**

PIC	***Legal name***
999978627	*JOHANNES GUTENBERG-UNIVERSITAT MAINZ*

Short name: JGU MAINZ

Address of the organisation

Street SAARSTRASSE 21

Town MAINZ

Postcode 55099

Country Germany

Webpage www.uni-mainz.de

Legal Status of your organisation

Research and Innovation legal statuses

Public body ....yes

Non-profit ....yes

International organisation ....no

International organisation of European interest ......no

Secondary or Higher education establishment ....yes

Research organisation ....yes

Legal person .... yes

Enterprise Data

SME self-declared status....................................17/07/2014 - no

SME self-assessment .... unknown

SME validation sme... unknown

Based on the above details of the Beneficiary Registry the organisation is not an SME (small- and medium-sized enterprise) for the call.

Department(s) carrying out the proposed work

Department 1

Department name	Data Center	□ not applicable
	□ Same as organisation address	
Street	Anselm-Franz-von-Bentzel-Weg 12	
Town	Mainz	
Postcode	55128	
Country	Germany	

Dependencies with other proposal participants

Character of dependence	***Participant***	

Proposal ID	**800929**	*Acronym*	**DIO**	*Short name*	**JGU MAINZ**

Person in charge of the proposal

The name and e-mail of contact persons are read-only in the administrative form, only additional details can be edited here. To give access rights and basic contact details of contact persons, please go back to Step 4 of the submission wizard and save the changes.

Title	Prof.	Sex	● Male ○ Female
First name	**André**	Last name	**Brinkmann**
E-Mail	**brinkman@uni-mainz.de**		
Position in org.	Head of Data Center and of Efficient Computing and Storage Group		
Department	Data Center	□ Same as organisation	
	□ Same as organisation address		
Street	Anselm-Franz-von-Bentzel-Weg 12		
Town	Mainz	Post code	55128
Country	Germany		
Website	https://research.zdv.uni-mainz.de/people/andre-brinkmann/		
Phone 1	+49 6131 3926390	Phone 2	*+xxx xxxxxxxxx*
Fax	*+xxx xxxxxxxxx*		

Other contact persons

First Name	***Last Name***	***E-mail***	***Phone***
Julia	Doré	eu-office@uni-mainz.de	+49 6131 3926865

Proposal ID **800929** *Acronym* **DIO** *Short name* **ICCS**

PIC	***Legal name***
999654356	*INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS*

Short name: ICCS

Address of the organisation

Street	Patission Str. 42
Town	ATHINA
Postcode	10682
Country	Greece
Webpage	www.iccs.gr

Legal Status of your organisation

Research and Innovation legal statuses

Public body ....yes

Legal person .... yes

Non-profit ....yes

International organisation ....no

International organisation of European interest ......no

Secondary or Higher education establishment ....no

Research organisation ....yes

Enterprise Data

SME self-declared status..................................07/10/2008 - no

SME self-assessment .... unknown

SME validation sme..07/10/2008 - no

Based on the above details of the Beneficiary Registry the organisation is not an SME (small- and medium-sized enterprise) for the call.

Department(s) carrying out the proposed work

Department 1

Department name	Microlab, ECE, ICCS	□ not applicable
	□ Same as organisation address	
Street	9 Heroon Polytechneiou, Zografou Campus	
Town	Athens	
Postcode	15780	
Country	Greece	

Dependencies with other proposal participants

Character of dependence	***Participant***	

Person in charge of the proposal

The name and e-mail of contact persons are read-only in the administrative form, only additional details can be edited here. To give access rights and basic contact details of contact persons, please go back to Step 4 of the submission wizard and save the changes.

Title: Prof.

Sex: ● Male ○ Female

First name: **Dimitrios**

Last name: **Soudris**

E-Mail: **dsoudris@microlab.ntua.gr**

Position in org.: Associate Professor

Department: Microlab, ECE, ICCS

□ Same as organisation

□ Same as organisation address

Street: 9 Heroon Polytechneiou, Zografou Campus

Town: Athens

Post code: 15780

Country: Greece

Website: http://www.microlab.ntua.gr/~dsoudris/

Phone 1: +30 210 7724270

Phone 2: +30 210 7723653

Fax: +30 210 7722428

PIC	***Legal name***
913748731	*CYBELETECH SAS*

Short name: Cybeletech

Address of the organisation

Street	2 rue de la piquetterie
Town	Bruyères le Chatel
Postcode	91680
Country	France
Webpage	www.cybeletech.com

Legal Status of your organisation

Research and Innovation legal statuses

Public body ....no

Legal person .... yes

Non-profit ....no

International organisation ....no

International organisation of European interest ......no

Secondary or Higher education establishment ....no

Research organisation ....no

Enterprise Data

SME self-declared status..................................13/02/2017 - yes

SME self-assessment .... unknown

SME validation sme.. unknown

Based on the above details of the Beneficiary Registry the organisation is an SME (small- and medium-sized enterprise) for the call.

Department(s) carrying out the proposed work

No department involved

Department name ☒ not applicable

□ Same as organisation address

Street *Please enter street name and number.*

Town

Postcode

Country

Dependencies with other proposal participants

Character of dependence	***Participant***	

Proposal ID **800929** *Acronym* **DIO** *Short name* **Cybeletech**

Person in charge of the proposal

The name and e-mail of contact persons are read-only in the administrative form, only additional details can be edited here. To give access rights and basic contact details of contact persons, please go back to Step 4 of the submission wizard and save the changes.

Title	Dr.	Sex	● Male ○ Female
First name	**Denis**	Last name	**Wouters**
E-Mail	**denis.wouters@cybeletech.com**		
Position in org.	Chief Technology Officer		
Department	*Please indicate the department of the Contact Point above in the organisation*	□ Same as organisation	
	□ Same as organisation address		
Street	LAB'O, 1 Avenue du Champ de Mars		
Town	ORLEANS CEDEX	Post code	45074
Country	France		
Website	http://www.cybeletech.com		
Phone 1	33 972104744	Phone 2	*+xxx xxxxxxxxx*
Fax	*+xxx xxxxxxxxx*		

3 - Budget for the proposal

No	Participant	Country	(A) Direct personnel costs/€ ?	(B) Other direct costs/€ ?	(C) Direct costs of sub-contracting/€ ?	(D) Direct costs of providing financial support to third parties/€ ?	(E) Costs of inkind contributions not used on the beneficiary's premises/€ ?	(F) Indirect Costs / € (=0.25(A+B-E)) ?	(G) Special unit costs covering direct & indirect costs / € ?	(H) Total estimated eligible costs / € (=A+B+C+D+F+G) ?	(I) Reimburse-ment rate (%) ?	(J) Max.EU Contribution / € (=H*I) ?	(K) Requested EU Contribution/ € ?
1	Forth	EL	566400	82000	0	0	0	162100,00	0	810500,00	100	810500,00	810500,00
2	Bsc	ES	460000	20625	0	0	0	120156,25	0	600781,25	100	600781,25	600781,25
3	Stfc	UK	447930	55000	0	0	0	125732,50	0	628662,50	100	628662,50	628662,50
4	Bull	FR	454230	53000	0	0	0	126807,50	0	634037,50	100	634037,50	634037,50
5	Jgu Mainz	DE	371000	24000	0	0	0	98750,00	0	493750,00	100	493750,00	493750,00
6	Iccs	EL	348400	43000	0	0	0	97850,00	0	489250,00	100	489250,00	489250,00
7	Cybeletech	FR	256250	15000	0	0	0	67812,50	0	339062,50	100	339062,50	339062,50
	Total		2904210	292625	0	0	0	799208,75	0	3996043,75		3996043,75	3996043,75

4 - Ethics issues table

1. HUMAN EMBRYOS/FOETUSES		Page
Does your research involve Human Embryonic Stem Cells (hESCs)?	○ Yes ● No	
Does your research involve the use of human embryos?	○ Yes ● No	
Does your research involve the use of human foetal tissues / cells?	○ Yes ● No	
2. HUMANS		Page
Does your research involve human participants?	○ Yes ● No	
Does your research involve physical interventions on the study participants?	○ Yes ● No	
3. HUMAN CELLS / TISSUES		Page
Does your research involve human cells or tissues (other than from Human Embryos/ Foetuses, i.e. section 1)?	○ Yes ● No	
4. PERSONAL DATA		Page
Does your research involve personal data collection and/or processing?	○ Yes ● No	
Does your research involve further processing of previously collected personal data (secondary use)?	○ Yes ● No	
5. ANIMALS		Page
Does your research involve animals?	○ Yes ● No	
6. THIRD COUNTRIES		Page
In case non-EU countries are involved, do the research related activities undertaken in these countries raise potential ethics issues?	○ Yes ● No	
Do you plan to use local resources (e.g. animal and/or human tissue samples, genetic material, live animals, human remains, materials of historical value, endangered fauna or flora samples, etc.)?	○ Yes ● No	
Do you plan to import any material - including personal data - from non-EU countries into the EU?	○ Yes ● No	
Do you plan to export any material - including personal data - from the EU to non-EU countries?	○ Yes ● No	
In case your research involves low and/or lower middle income countries, are any benefits-sharing actions planned?	○ Yes ● No	
Could the situation in the country put the individuals taking part in the research at risk?	○ Yes ● No	

7. ENVIRONMENT & HEALTH and SAFETY		Page
Does your research involve the use of elements that may cause harm to the environment, to animals or plants?	○ Yes ● No	
Does your research deal with endangered fauna and/or flora and/or protected areas?	○ Yes ● No	
Does your research involve the use of elements that may cause harm to humans, including research staff?	○ Yes ● No	
8. DUAL USE		Page
Does your research involve dual-use items in the sense of Regulation 428/2009, or other items for which an authorisation is required?	○ Yes ● No	
9. EXCLUSIVE FOCUS ON CIVIL APPLICATIONS		Page
Could your research raise concerns regarding the exclusive focus on civil applications?	○ Yes ● No	
10. MISUSE		Page
Does your research have the potential for misuse of research results?	○ Yes ● No	
11. OTHER ETHICS ISSUES		Page
Are there any other ethics issues that should be taken into consideration? Please specify	○ Yes ● No	

I confirm that I have taken into account all ethics issues described above and that, if any ethics issues apply, I will complete the ethics self-assessment and attach the required documents. ☒

How to Complete your Ethics Self-Assessment

Proposal ID **800929** *Acronym* **DIO**

5 - Call specific questions

Extended Open Research Data Pilot in Horizon 2020

If selected, applicants will by default participate in the Pilot on Open Research Data in Horizon 2020[1] , which aims to improve and maximise access to and re-use of research data generated by actions.

However, participation in the Pilot is flexible in the sense that it does not mean that all research data needs to be open. After the action has started, participants will formulate a Data Management Plan (DMP), which should address the relevant aspects of making data FAIR – findable, accessible, interoperable and re-usable, including what data the project will generate, whether and how it will be made accessible for verification and re-use, and how it will be curated and preserved. Through this DMP projects can define certain datasets to remain closed according to the principle "as open as possible, as closed as necessary". A Data Management Plan does not have to be submitted at the proposal stage.

Furthermore, applicants also have the possibility to opt out of this Pilot completely at any stage (before or after the grant signature). In this case, applicants must indicate a reason for this choice (see options below).

Please note that participation in this Pilot does not constitute part of the evaluation process. Proposals will not be penalised for opting out.

We wish to opt out of the Pilot on Open Research Data in Horizon 2020.	○ Yes	● No

Further guidance on open access and research data management is available on the participant portal: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm and in general annex L of the Work Programme.

[1] *According to article 43.2 of Regulation (EU) No 1290/2013 of the European Parliament and of the Council, of 11 December 2013, laying down the rules for participation and dissemination in "Horizon 2020 - the Framework Programme for Research and Innovation (2014-2020)" and repealing Regulation (EC) No 1906/2006.*

COVER PAGE

DIO: High-Density I/O Path for Fast Storage Devices and Multi-Tier Hierarchies

Call Identifier: FETHPC-02-2017

Topic: Transition to Exascale Computing

Subtopic: (c) Exascale I/O and storage in the presence of multiple tiers of data storage

Types of action: RIA Research and Innovation action

List of Participants

Participant #	Participant Organisation Name	Part. Short Name	Country
1 (Coordinator)	Foundation for Research and Technology – Hellas	FORTH	Greece
2	Barcelona Supercomputing Center	BSC	Spain
3	Science and Technology Facilities Council	STFC	UK
4	BULL SAS	BULL	France
5	Johannes Gutenberg University Mainz	JGU	Germany
6	Institute of Communication and Computer Systems	ICCS	Greece
7	CybeleTech SAS	CYB	France

Table of Contents

1 Excellence

1.1 Objectives

Storage I/O in HPC systems is a main bottleneck today and more so in the future, especially with the current data growth rates. Today, HPC systems typically perform I/O using the general architecture shown in Figure 1. Storage devices, mostly Solid State Drives (SSDs) and Hard Disk Drives (HDDs), are placed in external Storage Area Networks (SAN), and are managed by parallel file systems. However, contemporary filesystem servers and clients cannot scale in number similarly to compute nodes due to the file-level protocols used. For this purpose, today, there is a set of I/O aggregator nodes (subset of the compute nodes) that receive all I/O requests from compute nodes and forward them (with numerous optimizations) to the filesystem. Compute nodes use I/O libraries to communicate with aggregator nodes.

This architecture, however, cannot cope with two essential technology trends:

(1) The increasing adoption of faster storage devices, such as SSDs and NVM that access data at microsecond-level latencies, as well as emerging storage technologies, such as SCM.

(2) The fact that future systems need to scale out to hundreds of thousands of compute nodes to cope with increasing problem sizes and datasets.

Adding fast devices to the SAN cannot provide the required level of performance and scalability because of the finite throughput between compute nodes and the SAN itself. Therefore, emerging device technologies need to be placed in compute nodes. However, the parallel file system is not able to manage local devices in compute nodes because it incurs high overhead compared to device latency. Furthermore, parallel filesystem metadata management does not scale.

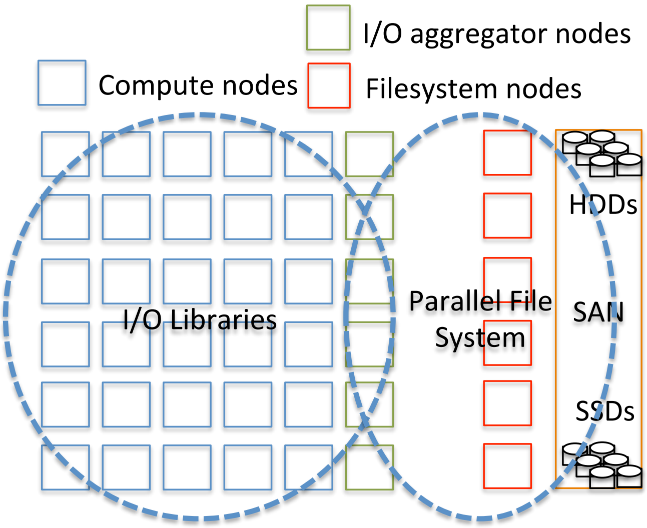


Figure 1: storage architecture of HPC systems.

DIO addresses these problems by enabling the seamless use of deep storage hierarchies, as follows:

(1) **It takes advantage of technology trends:** It allows fast storage devices to be placed locally in the compute nodes and the rack, at several tiers, depending on evolving device technology.

(2) **It proposes new technology for the I/O path:** It proposes a fundamentally novel software stack for the I/O path that manages these devices in an efficient manner, using principles of modern key-value store systems (WP2), efficient I/O abstractions and libraries (WP3), in-transit acceleration inline with the promising near-compute paradigm (WP5), and continuous monitoring and optimization for I/O performance (WP4).

(3) **It facilitates deployment:** It maintains the parallel filesystem as the *last tier* to facilitate deployment. It also provides the necessary glue for I/O libraries to run over the new storage system that manages fast storage tiers, and therefore, eliminates or minimizes application modifications.

DIO will design, prototype, and evaluate the necessary systems software stack to allow future HPC systems to benefit from new storage devices and deep storage hierarchies in a manner that facilitates deployment with legacy systems and applications.

*Overall, DIO has the following objectives and corresponding quantifiable targets (*Table 1*):*

Table 1: Objectives and Quantifiable Targets of DIO.

Objective	Quantifiable Target
Technology targets	
1. DIO will use a novel approach to provide a global I/O address space via partitioning, placing, and indexing data on devices, based on multi-level write-optimized data structures that have emerged in the area of data analytics and have the ability to inherently adapt to different device technologies at each level/tier and achieve unprecedented rate of data movement between application memory and storage devices. **[WP2]**	*Demonstrate a running system at rack scale with at least four tiers of storage devices from available technologies (DRAM, SCM, NVMe, SSDs, HDDs), which will achieve a data transfer rate of 10 GB/s per server (about 160 GB/s per rack, assuming 16 fat servers per rack), which for 100 racks translates to 16 TB/s, or 1 PB/min, or 1 EB/day. This is more than 10x improvement over today's systems that move data at about 10 GB/s per rack.*
2. DIO will reduce or eliminate systems software overheads,	*Reduce by 10x CPU overhead per I/O operation*

(such as kernel, frequent cache lookups, network protocol processing) by using memory mapped I/O over fast storage devices. **[WP2]**	*that today is several tens to hundreds of microseconds (tens to hundreds of thousands of CPU cycles).*
3. DIO will provide both legacy API support, and novel APIs optimized for the I/O path acceleration. **[WP3]**	*Enable legacy applications to run unmodified by porting the popular MPI-IO, HDF5, NetCDF, XIOS libraries and allow applications written with new APIs to benefit from direct and low overhead access to data in fast tiers.*
4. DIO will take advantage of multi-tier storage and fast devices to provide efficient support for checkpointing. **[WP3]**	*Reduce checkpoint interval from hours to mins without affecting application performance.*
5. DIO will enable sharing, access, and prioritization at fine grain and on a per key-value pair basis. DIO will provide mechanisms and policies for allocating resources to concurrent applications and diverse workloads that exhibit high-degrees of parallelism and require dynamic data placement across tiers to achieve deterministic application QoS. **[WP4]**	*Allow resources in the I/O path to be shared across many concurrent jobs with individual QoS targets, eliminating sources of interference, and providing to applications a fair share of at least 50% of the total physical device throughput.*
6. DIO will seamlessly and efficiently integrate FPGA acceleration in the I/O path via a data-oriented API and appropriate accelerator design for near-device processing to reduce overheads of application-induced transformations. **[WP6, WP3]**	*Enable in-transit data processing using accelerators at line rate (as per Objective 1, at about 10 GB/s per server) while data is moved between memory and persistent storage.*
Demonstration and use of technology targets	
7. DIO will demonstrate benefits through a working prototype and the use of production workloads (Code_Saturne in multi-physics, NEMO in environmental modeling, and Cybele in agricultural production forecast). **[WP5, WP6]**	*Demonstrate the impact of DIO technology on a working rack-scale prototype with the full I/O stack, commercial storage devices, accelerators, and real applications by achieving:* • *Code_Saturne: Finer-grain interactive visualization at 10x larger dataset sizes.* • *NEMO: 10x larger model resolution.* • *Cybele: 10x larger regions during training and <2% forecast error for prediction.*
8. DIO will facilitate deployment and adoption (a) by examining important parameters for systems that aim at exascale performance **[WP6]** and (b) working with vendors to show how future systems can incorporate the proposed systems software stack in data-intensive HPC applications. **[WP7]**	*Determine using extrapolation how DIO will scale to 100s of racks and exascale performance, present a technology roadmap for the future adoption of the proposed I/O technology, and design 3-4 Proof-of-Concept installations in operational environments in partners after the project.*

1.1 Relation to the work programme

DIO is a proposal to the call "*FETHPC-02-2017 Transition to Exascale Computing",* subtopic *"(c)* Exascale I/O and storage in the presence of multiple tiers of data storage". DIO exhibits excellent relationship to the work programme both in terms of its goals but also the expertise of all project partners. The following table depicts the alignment of DIO to the work program and the specific call.

Proposals should address exascale I/O systems expected to have multiple tiers of data storage technologies, including non-volatile memory.
DIO inherently addresses the issues of multi-tier storage using different and diverse device technologies. It allows storage devices to be placed within compute nodes and provides a systems software stack that reduces or eliminates I/O overheads to allow applications to benefit from improved device throughput and latency. DIO's approach to use a multi-level key-value store and memory mapped I/O for data access provide the fundamental mechanisms for seamlessly integrating existing and emerging storage devices in the I/O path. DIO allows NVM memory to be used as tier-0 in each compute node and further explores the potential of byte-addressability for

metadata management in the I/O path via work on novel APIs and I/O libraries.
Fine grain data access prioritization of processes and applications sharing data in these tiers is one of the goals as well as prioritization applied to file/object creates/deletes.
DIO will allow fine grain data sharing and access prioritization at two levels: 1) Arbitrary sized key-value pairs and 2) key regions. Sharing and prioritization at these levels of the key-value store is made possible by the efficient indexing structure and associated metadata used by DIO. Additionally, DIO will enable placement of data on different tiers (and nodes), both implicit (driven by use) and explicit (via a high-level API). DIO will allow applications to generate large numbers of concurrent operations and will have the ability to schedule them implicitly or explicitly to satisfy application QoS, especially when concerning user-facing workflows.
Runtime layers should combine data replication with data layout transformations relevant for HPC, in order to meet the needs for improved performance and resiliency.
DIO will provide efficient (asynchronous) data replication over RDMA and will also examine approaches to improve device efficiency, compared to full data replication. In addition, DIO will allow data replicas to be transformed while being transferred in the I/O path between storage devices and application memory by using appropriate extensions to the I/O path API. DIO aims at making data transformations a first-class citizen in the I/O path by use of transparent acceleration, rather than performing all associated processing in the application address space, which typically results in significant overheads.
It is also desirable for the I/O subsystem to adaptively provide optimal performance or reliability especially in the presence of millions of processes simultaneously doing I/O.
DIO will include extensive mechanisms for resource allocation, QoS specification, monitoring, and adaptation. This is an important aspect of DIO that addresses interference induced by multiple applications using the same system resources. DIO's approach is end-to-end, because it starts from application requirements in QoS, translates them to the storage resources required, assigns and isolates the use of the resources to multiple, concurrent applications, and monitors application behavior to observe deviations from the desired QoS.
It is critical that programming system interoperability and standardized APIs are achieved.
DIO addresses this requirement in two ways. First, it allows applications and libraries to use customized APIs without having to incur the overhead of generic data management functions and expensive protocols, as is typically the case today. Second, DIO will implement popular libraries, such as MPI-IO, HDF5, NetCDF, and XIOS that will allow applications to run unmodified. In this respect, DIO is both friendly towards legacy applications but also provides the path to evolve application interfaces and semantics for storage I/O.
On the fly data management supporting data processing, taking into account multi-tiered storage and involving real time in situ/in transit processing should be addressed.
DIO addresses this requirement by providing the ability to (a) transparently place data to different tiers and nodes, (b) allow applications to explicitly place data to a prefered tier, (c) to inquire about the location of data so computation can be optimized, (d) to perform data-processing transformation while data are being moved between application memory and persistent storage using FPGA accelerators.

1.2 Concept and methodology

1.2.1 Concept

Storage I/O in modern HPC infrastructures and datacenters is a cornerstone for keeping up with data growth. Modern HPC and other data processing applications generate and consume large amounts of data to generate value for science, society, and industry. However, scaling storage I/O and keeping up with data growth has historically been a challenge and is today a main enabler for further improving computing infrastructures.

Today, there are two opportunities to address I/O issues: (a) device latency has improved dramatically with the emergence of SSDs and NVM and continues to improve, e.g. with emerging SCM and (b) we can use larger numbers of storage devices, especially NVM-grade devices, in compute nodes themselves. However, the current systems software stack in the I/O path cannot exploit the entire potential of these opportunities, because it exhibits high overheads, limited scalability, and limited extensibility.

The fundamental problem that needs to be addressed is the improvement of the efficiency of I/O, as overheads have shifted from devices to the systems software stack[1]: An I/O operation that takes (or will take) 1 microsecond to serve by a device, requires today 10s or 100s of microseconds processing at the host level. However, increasing I/O efficiency and data serving capabilities is a challenging problem, because of three main reasons:

- Device heterogeneity and multiple storage tiers
- Diverse workloads
- Software overhead in the I/O path
- Interfaces and abstractions

Device heterogeneity means that no single device technology is the best in all important dimensions: capacity (density), cost, throughput, IOPS, latency, data retention (reliability), and performance predictability. For instance, some devices achieve the best cost/GB (HDDs) and other devices the best cost/IOPS (FLASH-based). In addition, this diversity in device heterogeneity is projected to increase due to technology. Therefore, storage systems will employ multiple types of devices, typically in the form of storage tiers, which leads to high complexity in the architecture of the storage system. Furthermore, this complexity is exacerbated by the need to support diverse workloads, where one cannot optimize the storage system for one dimension of the complex parameter space. In fact, storage system complexity and workload diversity is emerging as an important problem due to the size of the underlying system, as we head towards exascale-level I/O. Complexity and diversity reach such levels that it becomes impossible for individual applications (and their programmers/users) to optimize statically, requiring dynamic adaptation at runtime from the system itself.

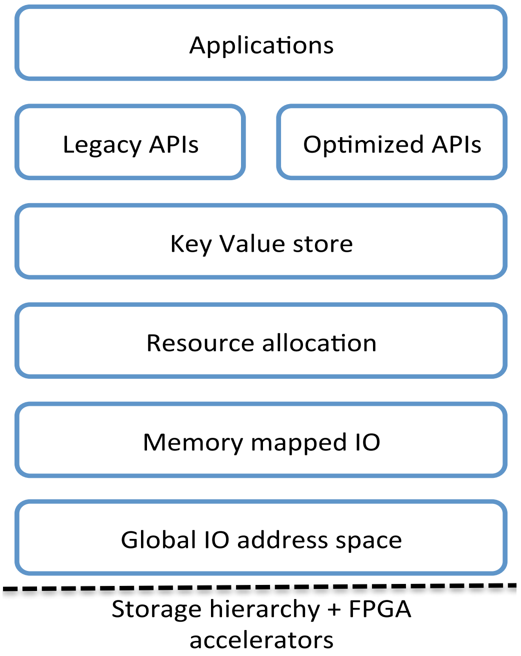


Figure 2: The proposed DIO software stack.

To manage complexity and diversity, the I/O path has evolved over time to include several techniques, such as caching, dynamic data placement, and logical to physical translation that typically happen in several layers in user space libraries and the operating system kernel. This layering and the need to support complex functions and metadata management, typically in kernel space, have led to high software overheads. For instance, I/O request processing in a modern system might involve hundreds of microseconds of processing, which eliminates any benefits from using fast storage devices, such as SSDs, NVMe, and NVM.

Finally, traditional I/O interfaces and abstractions at various levels in the I/O stack do not match modern technology capabilities, e.g. the use of accelerators for in-I/O-path data processing and transformation, or providing hints about what the application requires, or what the system should do.

Given these limitations, storage I/O comprises a bottleneck for many applications and services today that require data processing. When applications need to switch datasets between memory and persistent storage, they incur delays that currently inhibit scaling to larger datasets. DIO will overcome these limitations by designing, prototyping, and demonstrating the I/O software stack of Figure 2. Next, we discuss each aspect of the proposed technology in more detail.

POSITIONING WITH RESPECT TO TECHNOLOGY READINESS LEVELS

Table 2 presents the TRL's (Technology Readiness Levels) for all technological aspects applied to the project based on what has been described in previous sections.

Table 2: TRL for each DIO component/tecnology.

Component	Previous Work	Current TRL	Targeted TRL
Data access	Research on basic principles and various types of validation in lab for technology trends and approaches to consider	TRL3	TRL6
Data replication	Research on basic principles and technology trends that dictate different design approach for future platforms	TRL3	TRL6
Acceleration	Research on basic principles and technology trends that dictate different design approach for future platforms	TRL3	TRL6

[1] Luiz Barroso, Mike Marty, David Patterson, P. Ranganathan, "Attack of the Killer Microseconds", Commun. ACM, vol. 60, no. 4, April 2017.

APIs, Libraries	Research and experimental proof of concept	TRL3	TRL6
Sharing, isolation	Research and experimental proof of concept	TRL3	TRL6
Checkpointing	Research and experimental proof of concept	TRL3	TRL6
Application adaptation	Research how applications benefit from new storage capabilities	TRL3	TRL6
Integrated DIO Platform	Project partners have extensive expertise in multiple aspects of the proposed solution and application domains	TRL4	TRL6

Overall, DIO will deliver the system software stack for the I/O path that addresses fundamental limitations of current approaches and allows future systems to benefit from emerging storage device technologies and to keep up with data growth in modern HPC infrastructures.

LINKS TO OTHER PROJECTS AND TECHNOLOGIES

DIO will exploit the results produced by a series of previous or ongoing EU funded or National projects that focus on variety of aspects related to the DIO concept and objectives. Access to these projects is enabled through links with members of the DIO consortium and each partner's existing network. Table 3 outlines some of the projects that DIO has targeted as the most relevant for establishing strong and solid synergies between them.

Table 3: Relevant projects for establishing synergies.

Name	Results Relevant to DIO
H2020-NextGENIO	NextGENIO aims to offer a new system architecture based on NVM. The project is divided on hardware and software, and software has the objective to prepare tools, mechanisms and schedulers to use the hardware in several ways. "Does application A work better with traditional POSIX semantics or with objects?" This kind of questions is covered in the project. DIO will benefit from the knowledge obtained in several components as echoFS (an ephemeral transparent file system over NVM) and the different schedulers. BSC is a partner in NextGENIO, acting as WP leader of the system ware. DIO builds on the results of NextGENIO and aims to address issues in the software stack for I/O targeting fast devices and multi-tier hierarchies.
H2020-SAGE	The SAGE project builds a data centric infrastructure for handling extreme data in the Exascale era, centered on an object storage architecture, 'Percipient Storage', with the ability to accept and perform user defined computations integral to the storage system. STFC is a partner in SAGE, where it hosts and maintains the project's rack-scale prototype. DIO builds on SAGE and aims to address issues in the software stack for I/O targeting fast devices and multi-tier hierarchies.
H2020-ExaNode	ExaNoDe is developing a heterogeneous Multi-Chip-Module (MCM) combining the following key components: ARM-v8 computing architecture; 3-D integration (interposer) of System-on-Chips (SoC) for higher compute density combined with high-bandwidth, low-latency data communication interfaces; UNIMEM-based advanced memory schemes for high scalability. DIO is complementary to ExaNode as it targets the systems software I/O path.
H2020-ExaNest	ExaNeSt develops and prototypes solutions for Interconnection networks, storage device connectivity, and cooling for Exascale HPC to become feasible. ExaNeSt is developing rack-scale prototypes supporting the UNIMEM remote-memory model, and aims to port, tune and evaluate a diverse set of compute-oriented and data-management applications. FORTH is the coordinator of ExaNeSt and a key contributor in the WP focused on interconnects. DIO is complementary to ExaNest by building the systems software stack to deal with overheads in the I/O path, fast devices, and multi-tier hierarchies.
H2020 - VINEYARD	VINEYARD is a response to the observation that hardware accelerators, such as GPUs and FPGAs, expose a dual problem to a datacenter ecosystem: First, application developers find them hard to use, despite the fact that most of the computationally expensive routines have been already implemented by various High Performance Computing (HPC) experts; second they are hard to be shared across multiple tasks, either in intra-job level or across different jobs. VINEYARD aims to address related issues, making a step forward in how accelerators are accessed from host software. ICCS, FORTH, BULL, and STFC are partners in VINEYARD, while ICCS is the project coordinator. DIO will use VINEYARD mechanisms to integrate accelerators in the I/O path with the focus to perform simple but heavy data transformations on data in the more confined context of I/O operations and as data flow between application memory and storage devices. DIO is complementary to VINEYARD as it provides the systems software stack for the I/O path.
H2020-EuroExa	EuroEXA is a recently initiated project that builds on previous European high-performance computing projects and partnerships and brings together the focus of European industrial SMEs. The project results will be demonstrated on an integrated and operational prototype with a rich mix of key applications from across climate/weather, physics/energy and life-science/bio-informatics

	domains. FORTH is a partner, leading the systems software WP focusing on operating system and firmware issues for the three successive generations of the project's rack-scale testbeds. BSC is leading the applications software work-package focusing on porting and optimization of 14 HPC applications. STFC leads the WP on deploying and integration of the final fully supported novel system solution within the data center infrastructure. DIO is complementary to EuroEXA since it addresses fundamental issues in the I/O path. DIO will interact with EuroEXA to exchange information about technology trends, progress of development, and evaluation methodology.
ACTiCLOUD	ACTiCLOUD proposes a novel cloud computing architecture for drastically improved management of cloud resources, targeting a 1.5x increase in resource efficiency and more than 10x in scalability. By utilizing modest investments on hardware intelligence that enables true resource disaggregation between multiple servers, this project aims to progress current state-of-the-art in hypervisors and cloud management systems promoting holistic resource management at the rack scale and across distributed cloud sites. ACTiCLOUD tries to evolve the ecosystem around in-memory databases, an emerging component for a demanding class of applications that face difficulties in matching their resource requirements with state-of-the-art cloud offerings, with the goal to provide cost-efficient Database-as-a-Service (DBaaS) cloud platforms. ICCS is the coordinator of the ACTiCLOUD consortium. DIO's software stack will also have applicability in VM-based environments, therefore DIO will interact with ACTICLOUD on insights in the direction of converging HPC and Cloud domains, challenges, and opportunities.
Marie Curie ETN BigStorage	BigStorage is a Marie Curie ETN project that aims at training young researchers in the area of storage, taking into consideration emerging technology themes such as the use of NVM devices in the I/O path, data placement in a multi-tier storage environment, and the unification of HPC and Cloud storage. FORTH, BSC, and JGU are partners in BigStorage and will ensure that technology and results from DIO are used in the training aspects of BigStorage allowing early stage researchers of the project early access to research and associated prototypes.
H2020-M2DC	The main goal of M2DC is to develop a new class of low-power TCO-optimized appliances with built-in efficiency and dependability enhancements, easy to integrate with a broad ecosystem of management software and fully software-defined to enable optimization for a variety of future demanding applications in a cost-effective way. M2DC is complementary to DIO. The I/O path technology developed in DIO will be applicable also to M2DC servers in both single and multi-node setups to manage storage devices and data access.
DoE FastForward	The DOE FastForward initiatives (now FastForward-2) in the US aim to accelerate the development of extreme scale technologies. FastForward is a framework that provides funding for work in extreme scale, including work in the storage system and I/O path. DIO partners have contacts with organizations that participate in FastForward and will ensure links between efforts.
dRedBox	The dRedBox (FETHPC) project is shifting to employ pooled, disaggregated - instead of monolithic and tightly integrated components. The dReDBox proposition has the ambition to lead to significantly improved levels of utilization, scalability, reliability and power efficiency, both in conventional cloud and edge datacenters. BSC and FORTH are partners in this project, contributing in the development and evaluation of the project rack-scale prototype. DIO is complementary to dRedBox as it offers a software stack that can manage storage devices in the target architecture of dRedBox.
ESiWACE	ESiWACE is a Center of Excellence (CoE) that aims to substantially improve efficiency and productivity of numerical weather and climate simulation on high-performance computing platforms by supporting the end-to-end workflow of global Earth system modelling in HPC environment. NEMO is one of the central applications in the CoE and therefore synergies with DIO are strong. BSC, STFC, and BULL are partners in ESiWACE and they will ensure that DIO technology related to storage and I/O will be disseminated towards the community of ESiWACE so that potential synergies can be developed.

1.2.2 Methodology

DIO will use a new approach to provide data access and data management to HPC applications that generate, use, or manipulate large amounts of data. It will also implement a QoS aware resource management scheme that enables the sharing of the I/O path among multiple processes. Figure 3 shows at a high level the overall hardware architecture and Figure 4 the systems software stack in DIO. Figure 5 shows the internal components of DIO, which later on will be mapped to WPs.

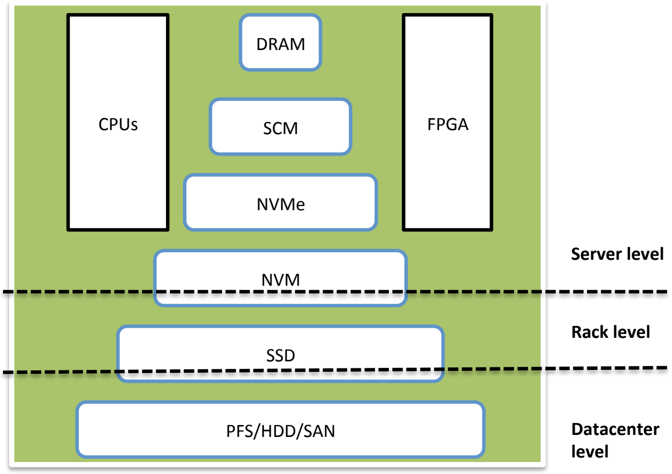


Figure 3: Conceptual hardware hierarchy in DIO. Device technologies at the boundaries can be placed on either side, depending on the envisioned application. Although rack-scale storage today is less popular due to the lack of appropriate interconnects, it is emerging as a new class with developments of new protocols for remote access, such as NVMe over Fabrics[2].

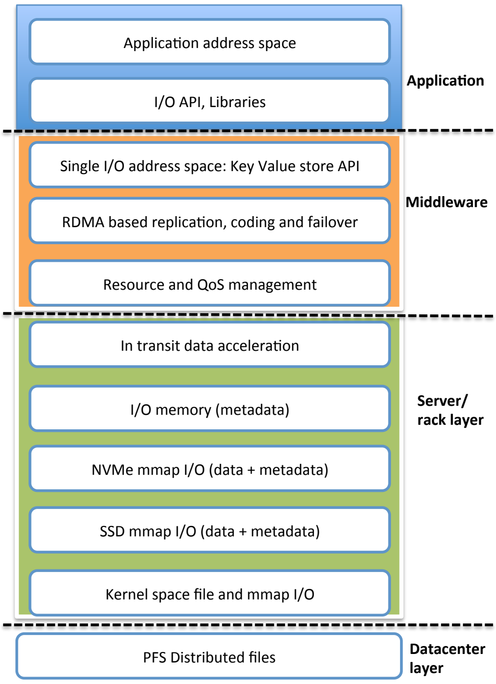

Application address space
I/O API, Libraries
Application
Single I/O address space: Key Value store API
RDMA based replication, coding and failover
Resource and QoS management
Middleware
In transit data acceleration
I/O memory (metadata)
NVMe mmap I/O (data + metadata)
SSD mmap I/O (data + metadata)
Kernel space file and mmap I/O
Server/
rack layer
PFS Distributed files
Datacenter
layer

Figure 4: DIO software stack.

The rest of this section is organized as follows: First, we discuss each important aspect of the I/O path in DIO and how the proposal will achieve its goals in more detail. Afterwards, we discuss how we are going to implement both legacy and novel APIs for data access and how we will enable checkpointing for fault tolerant execution. Then, we discuss DIO's solution to resource allocation and how DIO use FPGA accelerators to offload common, expensive operations. Finally, we discuss how we are going to run production level procurement benchmarks (UEABS) and three real applications from diverse domains: environmental modeling (NEMO/XIOS), a High Performance Data Analytics use case in agricultural production forecast (Cybele), and multi-physics (Code_Saturne).

DATA ACCESS AND METADATA MANAGEMENT

DIO proposes to replace traditional storage device access mechanisms, such as, files, objects and blocks with a new key-value store abstraction. Although the traditional mechanisms have been explored extensively and are in use today in most high-performance storage systems, mainly in the form of parallel file systems or scalable object storage systems, there are two main disadvantages of current approaches:

- They have been designed for slow hard disk drives and they are not appropriate for faster storage devices, such as SSDs, NVMe, and byte-addressable NVM, due to their overhead, including protocols for metadata management, user-kernel space crossings, etc. Recently, there have been efforts to adapt these systems to fast devices, e.g. by using SSD-caching, which however does not address the fundamental problems.
- The APIs they offer are not appropriate for viewing data as datasets that need to be accessed (generated, read, manipulated) but rather they offer a system-view of a sequence of bytes or blocks. Therefore, they lead to the use of additional layers for presenting application-friendly views of the data, with additional overheads, especially at large scale. Additionally, the data abstractions offered by these systems make it hard to optimize their structure on emerging devices and also to incorporate acceleration in the I/O path.

DIO will address these issues with a different approach to data access: We will use an indexing and metadata management system based on modern key-value stores that aim to provide efficient access to large amounts of data. Key-value stores[3,4,5] have recently been proposed and are used extensively to handle large datasets. The abstraction offered by key-value stores is typically a dictionary abstraction that supports simple operations: get(), put(), scan(), and delete(), with possible extensions for versioned items and management operations, key-range split and merge. Although this is a simple abstraction over stored data, it has proven to be extremely powerful and convenient for building modern services and applications over large amounts of data.

[2] http://www.nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf

[3] GOOGLE. Leveldb. http://leveldb.org/. Accessed: May 23, 2016.

[4] FACEBOOK. Rocksdb. http://rocksdb.org/. Accessed: May 23, 2016.

[5] A. Papagiannis, G. Saloustros, P. González-Férez, and A. Bilas. 2016. Tucana: design and implementation of a fast and efficient scale-up key-value store. USENIX ATC '16. USENIX Association, Berkeley, CA, USA, 537-550.

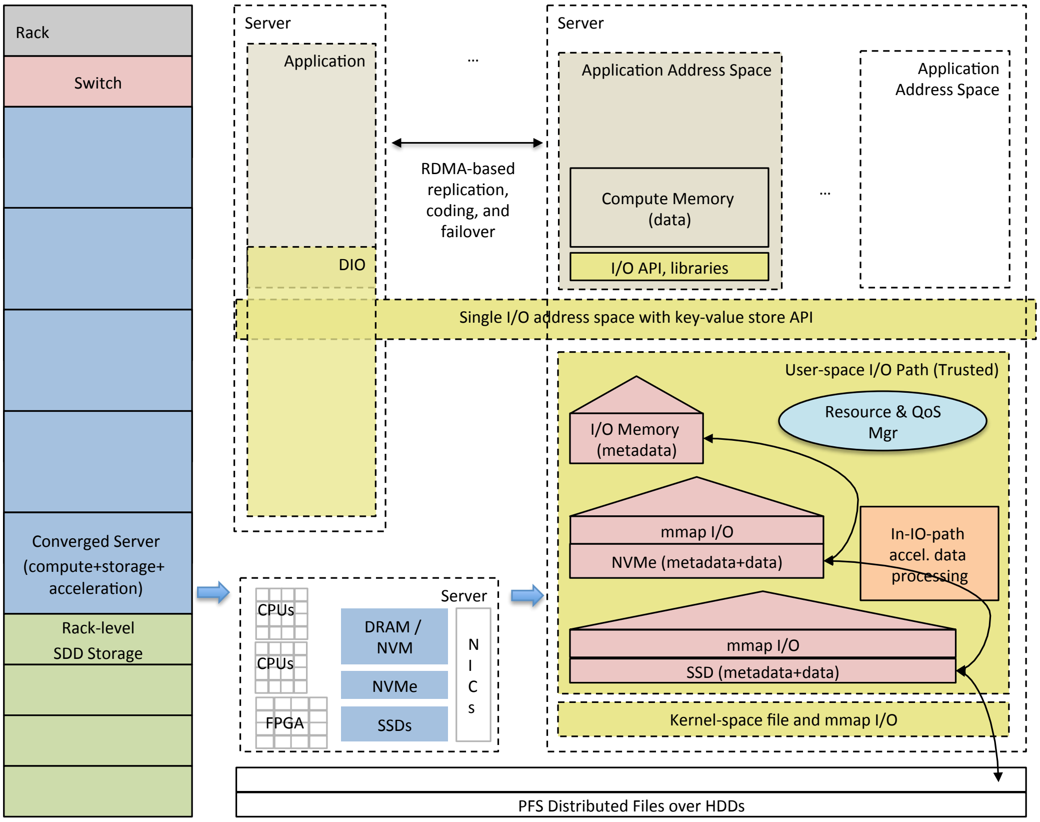

Rack
Switch
Converged Server (compute+storage+ acceleration)
Rack-level SDD Storage
Server
Application
DIO
...
RDMA-based replication, coding, and failover
Server
CPUs
CPUs
FPGA
DRAM / NVM
NVMe
SSDs
N I C s
Server
Application Address Space
Compute Memory (data)
I/O API, libraries
...
Application Address Space
Single I/O address space with key-value store API
User-space I/O Path (Trusted)
I/O Memory (metadata)
Resource & QoS Mgr
mmap I/O
NVMe (metadata+data)
In-IO-path accel. data processing
mmap I/O
SSD (metadata+data)
Kernel-space file and mmap I/O
PFS Distributed Files over HDDs

Figure 5: Overview of the DIO components.

Key-value stores are typically able to handle bursts of writes well (e.g. as in checkpointing) by managing storage devices in a manner that results in mostly large I/Os. To achieve this, they use new techniques for storing data and metadata, amortizing write operations, merging updates to data, organizing data on devices and keeping metadata in memory. Therefore, key-value stores present an important opportunity for scaling I/O and improving the performance not only of common-path (e.g. read, write) but also management (e.g. create, delete) operations.

DIO Approach

State-of-the-art key-value (KV) stores typically use a Log-Structured Merge (LSM)-tree data structure[6] at their core for indexing. LSM-trees organize key-value pairs in multiple levels of large, sorted buckets, whose size increases at higher levels. Additionally, small amounts of search metadata, such as Bloom filters, are used to accelerate scan and get operations. This organization has two advantages. First, it requires a small amount of search metadata because containers are sorted and therefore, practically all I/O operations generated are related to data items (keys and values). Second, due to bucket sizes, I/Os can be large, up to several MB each, resulting in optimal device performance. LSM-trees are a write-optimized structure that is a good fit for HDDs where there is a large difference in performance between random and sequential accesses.

LSM-tree operation, however, in key value stores results in two major performance barriers. The first barrier is the overheads of compaction operations. In order to keep large data buckets sorted, LSM-trees perform compactions, which (a) incur high CPU overhead and (b) result in I/O amplification for reads and writes. The second performance barrier of today's key-value stores has to do with metadata and data access that requires traversing some form of an index. In all persistent key-value stores, the index includes pointers to data items (values) in the storage address space. During system operation, part of the index and data are cached in memory. When traversing the index to serve an operation, there is a need to translate storage pointers to pointers in memory. This leads to frequent cache lookups that cannot be avoided easily. Essentially, the cache serves as a mechanism to translate pointers from the storage to the memory address space. Most key-value stores today follow this caching approach. This allows the key-value store to also control the size and timing of I/O operations between the memory cache and the storage devices, as well as the cache policy. This approach to caching leads to excessive overheads as storage devices become faster. Even, with traditional storage systems, when all data and metadata fit in memory and all lookup operations hit, managing this cache (library calls for lookups) requires about one-third of the index CPU

[6] O'NEIL, P., CHENG, E., GAWLICK, D., AND O'NEIL, E. The log-structured merge-tree (lsm-tree). Acta Informatica 33, 4 (1996), 351–385.

cycles. DIO’s approach eliminates overheads associated with compactions by maintaining a different type of index over the data and overheads associated with data access by using memory mapped I/O.

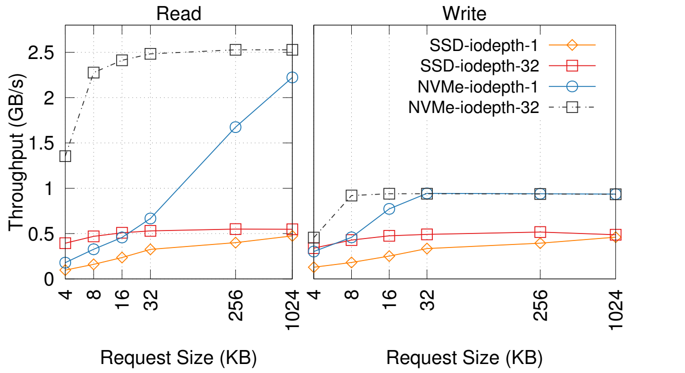


Figure 6: Throughput vs block size for SSD (Samsung 850 Pro 256GB) and NVMe (Samsung 950 Pro NVMe 256GB) devices, measured with the FIO[7] benchmark.

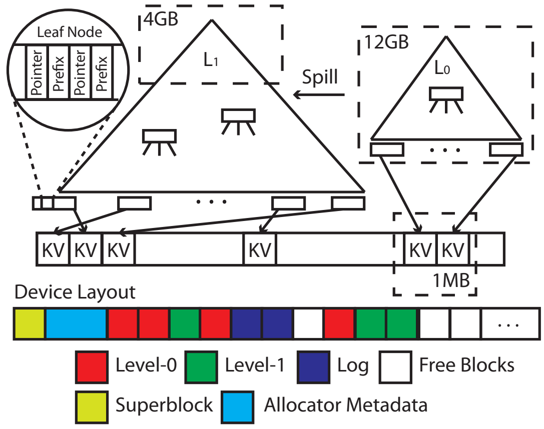


Figure 7: DIO main structures showing two levels of indexes, the key-value log, and device layout. Dashed rectangles indicate in-memory structures.

DIO will design and build a key-value store that is appropriate for fast storage devices and exhibits low overheads by using a new approach to data organization and indexing over fast devices, essentially trading (reducing) CPU overhead with device randomness and taking advantage of new storage technologies, as follows:

- DIO will take advantage of fast storage devices, such as SSDs and NVM to increase data serving density by reducing I/O amplification and CPU overhead. Figure 6 shows that modern SSD and NVMe devices are able to offer high performance, under high concurrency, even for small request sizes. Therefore, they present new opportunities for re-designing key-value stores and trading randomness for host overhead that has become the main I/O bottleneck.
- DIO will use a write optimized data structure that is organized in N levels, similar to LSM trees, where each level i acts as a buffer for the next level i+1. To reduce amplification, DIO will not operate on sorted buffers, but instead, it will maintain a B-tree index within each level. As a result, this approach generates smaller I/O requests in favor of reduced I/O amplification and CPU overhead. Figure 7 shows two levels of indexes, the key-value log, and device layout. Dashed rectangles include portions of the data structures that are kept in memory via mmap.
- DIO will also use an alternative approach based on mmap. Memory mapped I/O and mmap uses a single address space for both memory and storage and virtual memory protection to determine the location (memory or storage) of an item. This eliminates the need for pointer translation at the expense of page faults. We note that pointer translation occurs during index operations regardless of whether items are in memory or not, whereas page faults occur only when items are not in memory. The use of mmap will also allow DIO to use a single allocator for memory and device space management. Additionally, mmap eliminates data copies between kernel and user space. Given the presence of fast storage devices, such as NVM, the overhead of page faults is expected to be small and in the less common path.

DATA RECOVERY

Data recovery is the ability of a storage system to sustain any transient component failure without losing data. In storage systems, the focus is on storage devices that host data. Traditionally, storage systems (file or object systems) rely on some type of a log, typically in the form of a Write-Ahead-Log (WAL), to recover from such failures. Data is first written to the log and then subsequently transferred, in-place, to the actual device blocks that store the data.

DIO will use a Copy-on-Write (CoW) approach for persistence instead of a Write-Ahead-Log (WAL). WAL produces sequential write I/Os at the expense of doubling the amount of writes (in the log and later in place). CoW, on the other hand, performs only the necessary writes, however, it generates a more random I/O pattern. Therefore, although a WAL is more appropriate for HDDs, CoW has more potential for fast devices. The use of CoW is also motivated by three additional reasons; (a) It is amenable to supporting versioning. (b) It allows instantaneous recovery, without the need to redo or undo a log. (c) It helps increase concurrency by avoiding lock synchronization for different versions of each data item, as we discuss in the next subsection.

[7] https://github.com/axboe/fio

DIO will show that the Copy-on-Write approach is most appropriate going forward, as device technology evolves, and device performance approaches memory performance. Although the use of a CoW approach requires extensive redesign of the I/O path, this is in-line with the spirit and effort in DIO since we will provide a new design and implementation of the I/O path for modern devices, servers, and large datasets.

DEALING WITH TIERS IN DEEP STORAGE HIERARCHIES

Dealing with multiple storage tiers and deep storage hierarchies is one of the fundamental design principles behind DIO. There are three problems with managing emerging deep storage hierarchies: scalability, overhead, and data placement.

Scalability to large numbers of nodes (and devices):

Existing parallel file systems that manage storage devices, are not designed to manage the number of I/O nodes that is required by future systems. As storage devices will be placed in all or most of the compute nodes, there is a need to scale to hundreds of thousands of nodes. DIO takes advantage of fundamental benefits of key-value stores and proposes the design and use of a scale-out key-value store that operates over large numbers of devices and nodes, by using decoupled metadata across nodes and reducing or eliminating the need for persistent metadata.

Key-value stores (and DIO) are able to scale-out based on their design because each key-range is handled by a single node. Therefore, multiple accesses to the key-range are handled at a single point allowing for efficient handling of ordering and recovery of individual key ranges. Scalability is supported by the fact that an address space can be divided in large numbers of key-ranges and distributed to 100s of thousands of nodes.

DIO will rely on this fundamental aspect of key value stores to offer a single, partitioned address space. The address space can be partitioned to arbitrary key-value ranges, which then will be distributed dynamically to participating nodes, each node requiring only local operations for maintaining order and recovery in the common case. Additionally, DIO will allow key-ranges to be reconfigured based on application needs in terms of:

(a) Split a key-range to two key-ranges.
(b) Merge two consecutive key-ranges to a single key-range.
(c) Transfer a key-range from one node to another.

These mechanisms are essential for matching evolving application and workload requirements. Currently, such mechanisms exist, however, they incur high costs and more importantly they require application interruption during region reorganization. DIO's approach to use indexing at its core, allows it to perform split and merge operations without any data movement on devices and to provide transfers across nodes at the same time as data is being (re)placed on devices avoiding significant costs of today's systems and approaches.

Overheads

Fast storage devices, with I/O latencies at the microsecond level, require new approaches to device access and management so that system software overheads are reduced. For instance, most storage systems today live in the OS kernel and require user-kernel crossings that will dominate future device access times. DIO performs all I/O via memory mappings, eliminating the need for expensive lookups, not only in the kernel, but also in user space. Instead, it uses the virtual memory mechanism to implicitly identify I/O items that are located in memory. Therefore, DIO addresses overhead issues with fast storage devices.

Data placement

DIO will enable both coarse- and fine-grain data placement in different device tiers. For coarse data placement, DIO will allow creating multiple namespaces, e.g. over different device tiers and allowing the system profiler and scheduler to move data (key-value ranges) between namespaces and therefore, tiers. Although coarse-grain data placement is in many cases advantageous, in some cases it results in significant waste in space (and cost for expensive devices), e.g. when a single item (key) needs to be moved and by necessity carries with it a full key-region.

For fine-grain data placement DIO proposes an innovation at the key-value store level that will allow key-value pairs to be promoted from higher (slower) KV store levels (level(n)) to lower (faster) ones (level(n-1)). Data access methods designed to support multi-level structures in key-value stores are aware of different tiers. For instance, each level L_i of a key-value store can be placed on a different device technology, allowing for the system to adapt accesses to L_i to the properties of the corresponding device technology used in tier i. However, the important aspect that is missing today is the ability to tier-up (promote) part of the data from L_i to L_{i-1}. In DIO we will explore mechanisms and policies that will achieve this based on data use.

In terms of placement all data will eventually be stored in the last storage tier that is managed by the parallel file system as regular files. This is important for facilitating deployment and ensuring robustness. However, all performance aspects of the system will be managed directly by DIO over the fast storage tiers and devices. The dictionary API of DIO allows higher system layers (libraries and applications) to use simple mechanisms, such as tags and key-ranges to specify portions of the data (specific datasets) that should be placed in specific tiers. DIO will design and implement the corresponding APIs and data movement mechanisms allowing HPC (and cloud) applications to explicitly place data in specific tiers.

DATA AVAILABILITY

Data availability is the ability of a storage system to provide clients with access to data at all times. As a working assumption and at a high level, the main components that can fail and any approach needs to consider are: network paths (NICs, switches, links), servers (CPU, memory, power supply), and storage devices. To provide data availability, most systems use replication across nodes rather than within nodes. Therefore, every data item is replicated to K other nodes (and storage devices). This guarantees that data is available even if K-1 paths (any component in a path) fail. Data replication, that is the most popular approach, has two important drawbacks.

- Replication for writes slows down write operations. Every write needs to be replicated one or more times before a write operation completes. Unfortunately, in the presence of bursts of multiple write operations, replica update overheads dominate the write path, even in the best case, where one copy of the data is local to the node performing the write.
- Replication consumes equal space (K times) on each device involved. Although this has not been an issue with low-cost, high-density storage devices, such as HDDs, it is of paramount importance with expensive, fast devices, such as SSDs and NVMe today, and upcoming byte-addressable NVM devices in the near future. Such devices have a high cost and it is not realistic that their size can be doubled (or worse K-plied), given the system and data scales that are envisioned in future systems.

DIO will address these issues as follows:

RDMA can be used among compute nodes when exchanging I/O buffers. RDMA is already used in HPC systems for communication purposes, so our purpose is to integrate the I/O path with RDMA communication in an efficient manner. Unlike common storage systems, DIO has an advantage with the RDMA usage. In most approaches today, RDMA usage incurs overheads, because of required data copies between kernel and RDMA buffers. Instead, DIO eliminates those extra data copies because of its memory mapped design.

DIO will introduce asynchronous replication coupled with checkpointing semantics, in the sense that we will complete writes as soon as a single copy is stable and then asynchronously create the second copy. Although asynchronous replication does not delay writes, it creates a window of vulnerability. In HPC systems performing frequent checkpointing this window can be dealt with by keeping "alive" one more (previous) checkpoints and consider the current checkpoint stable only when all data has been replicated (and the window of vulnerability is closed). In this manner, the system can recover to a previous recovery point at the benefit of much faster writes. In addition, this approach allows us to choose as destination for replicas devices of a slower tier, which increases further the window, however, significantly reduces the cost for replication (in terms of space). This approach of co-designing replication and checkpointing semantics can have a significant benefit for multi-tier storage systems.

Third, DIO will examine the use of lightweight erasure coding techniques and protocols that have been proposed recently[8,9], instead of replication, for reducing replica space on fast (and expensive) devices. Erasure coding has been used extensively in low-cost (e.g. archival) systems where overhead is not a main concern. However, optimal erasure coding is generally an expensive approach in terms of both CPU and memory overheads. DIO will explore in its replication scheme the sustainability of erasure coding protocols for tier-0 devices to reduce space usage.

CHECKPOINTING

While executing in today's large and sophisticated HPC systems, large-scale jobs face an increasing risk of encountering system failures during their run. Checkpointing is a common operation used for recovering the global state[10] of long-running HPC applications in case of failure. Checkpointing places a high load on parallel file systems and a number of studies[11,12,13] estimate that it is expected to consume over 50% of total application

[8] Heng Zhang et al.. 2016. Efficient and available in-memory KV-store with hybrid erasure coding and replication. In *Proceedings of the 14th Usenix Conference on File and Storage Technologies* (FAST'16). USENIX Association, Berkeley, CA, USA, 167-180.

[9] G. J. Akash, et al. 2017. RAPID: A Fast Data Update Protocol in Erasure Coded Storage Systems for Big Data. In *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*(CCGrid '17). IEEE Press, Piscataway, NJ, USA, 890-897.

[10] Chandy, K. M., & Lamport, L. (February 1985). Distributed snapshots: Determining global states of distributed systems. *ACM Transactions on Computer Systems , 3* (1), 63-75.

[11] Oldfield, R. A., Arunagiri, S., Teller, P. J., Seelam, S., Varela, M. R., Riesen, R., et al. (2007). Modeling the Impact of Checkpoints on Next-Generation Systems. *Proceedings of the 24th IEEE Conference on Mass Storage Systems and Technologies (MSST'07).* Washington, DC, USA.

[12] I. R. Philp, Software failures and the road to a Petaflop machine, In Proceedings of 1st workshop on High Performance Computing Reliability Issues (HPCRI), 2005.

execution time in next-generation systems due to the size of application memory and the speed of the I/O path. *Achieving highly scalable checkpointing appropriate for exascale systems over emerging storage hierarchies is a key focus of DIO.*

Scalable multi-level checkpointing is an emerging methodology that alleviates the performance overhead of checkpointing on parallel file systems by combining frequent, lightweight checkpoints to handle common failure modes (such as one node failing temporarily) with less frequent, more expensive checkpoints for less common, but severe failures (all compute nodes failing). Faster checkpointing utilizes local storage, such as RAM (non-persistent), NVMe, SDD, or disk, and applies cross-node redundancy schemes, while the slowest but more complete and most resilient level writes to the parallel file system to withstand an entire system failure. Checkpoint files are occasionally flushed from cache to the parallel file system. Standard approaches to checkpointing that do not apply multi-level checkpointing are not expected to scale with the projected requirements of exascale systems.

There are two types of multi-level checkpointing, depending on the modification level of running applications. On the one hand, there is the transparent, system-level checkpointing, which has the advantage that applications communicating via MPI need not be modified; however, it has the significant drawback that it requires to collect unnecessarily large amounts of state for each applications, which results in high space and performance overhead. The main alternative approach is application-level checkpointing, which exhibits lower overhead and allows applications a higher degree of control over state management and parallel I/O. Applications typically checkpoint their state using explicit parallel I/O periodically to a single file through an I/O implementation (e.g., ROMIO[14]); they also use I/O interfaces to portable data formats such as NetCDF, parallel NetCDF, or HDF5, which in turn utilize MPI-IO over parallel file systems such as GPFS, Lustre, and OrangeFS; finally some applications adopt a simplistic "1 POSIX file per process" approach where each process checkpoints its state to a single file through a serial file I/O interface.

While significant prior work has focused on achieving high I/O rates with application-specific checkpointing in large-scale HPC environments, current approaches are projected to exhibit scalability problems in exascale systems, due to the standard practice of using a set of aggregator nodes to perform I/O towards a parallel file system or to excessive metadata overhead in the 1-file per process approach. Additionally, while POSIX is a standard interface today, most HPC applications (including those evaluated in DIO) use a parallel I/O library to perform I/O. POSIX I/O has unnecessarily strict semantics and requires locking mechanisms that inhibit the level of scalability necessary in exascale computing. *Thus, offering applications with access to checkpointing and storage in general via popular I/O libraries rather than directly via POSIX I/O is expected to provide coverage for the vast majority of existing HPC applications while being in line with prevailing trends towards the exascale era.*

A major goal of DIO *is to develop highly scalable multi-level checkpointing techniques that can support both transparent (system-level) and application-level checkpointing*. DIO will consider deduplication techniques[15] on checkpoints directly on the data path to support the range of HPC applications that use transparent checkpointing today. DIO is in position to achieve efficient deduplication, because its KV store can support content-defined addressing schemes. In addition, DIO's deviation from the POSIX standard will facilitate the implementation of application level checkpointing at exascale.

A key challenge of checkpointing implementations is to achieve asynchronous checkpointing so that applications can make progress during checkpointing operations. Existing approaches typically either perform synchronous checkpoints for simplicity or due to the lack of asynchronous APIs in I/O implementations, or require the maintenance of duplicate data structures, in-memory copying, and multi-threading schemes[16] to achieve asynchronous operation. *DIO can inherently achieve fully asynchronous checkpointing through the copy-on-write feature of its underlying key-value store, which allows low-overhead point-in-time snapshot operations on the underlying data sets across storage levels.*

Finally, with the growing use of specialized co-processors, such as field-programmable gate arrays (FPGAs) and graphics processing units (GPUs) as co-processors in HPC computations, checkpoint/restart over such schemes that include co-processor state becomes a necessity. A challenge coming up with accelerators is the difficulty to handle

[13] E. N. Elnozahy and J. S. Plank, Checkpointing for Petascale Systems: A look into the future of practical rollback-recovery, IEEE Transactions on Dependable and Secure Computing, 2004.

[14] Thakur, R., Gropp, W., & Lusk, a. E. (1999). On Implementing MPI-IO Portably and with High Performance. *Proceedings of Sixth Workshop on I/O in Parallel and Distributed Systems (IOPADS'99).* Atlanta, GA, USA.

[15] Kaiser, J., Gad, R., Süß, T., Padua, F., Nagel, L., Brinkmann, A.: *Deduplication Potential of HPC Applications' Checkpoints*. In Proceedings of the 2016 IEEE International Conference on Cluster Computing, CLUSTER 2016, Taipei, Taiwan, September 12-16, 2016.

[16] F. Shahzad, et al., Asynchronous Checkpointing by Dedicated Checkpoint Threads, Recent Advances in the Message Passing Interface: Proceedings of 19th European MPI Users' Group Meeting, EuroMPI 2012, Vienna, Austria, September 23-26, 2012.

co-processor computation state, typically managed independently from main-processor state. Previous work[17] has studied approaches to efficiently checkpointing the combined GPU-CPU memory state resident on machine nodes, addressing the end-to-end data movements required for checkpointing from GPU to storage, reducing checkpoint data movement cost seen by HPC applications. *DIO will undertake the extension of such unified checkpointing approaches to converged rack-scale server architectures featuring accelerators as co-processors and storage hierarchies including NVMe/SSD, a goal of key importance to the HPC community.*

LEGACY AND NEW STORAGE APIS

Typical storage systems today require that applications use a file or object API to access data. Both of these APIs follow a rather system-centric view of the world rather than a data-centric view. This means that applications need to adapt their internal data view to seemingly artificial APIs resulting in two significant sources of overhead: (a) The existing APIs generally incur high overhead because they involve a fair amount of processing, such as crossing user-kernel space boundaries, moving data, and performing various data management functions. (b) Applications require their own APIs and semantics, resulting in additional layers translating between these views, which require a lot of effort to design and build over existing storage systems. As a result, in many cases, applications do not customize existing APIs but rather accept associated overheads.

Two important aspects of DIO are: (a) The ability to offer new APIs to different applications within libraries or the application itself and to allow for easy customization, since there is no involvement of the kernel itself in the common data access path. (b) The ability to place data directly in the application address space without additional transfers and without the need to traverse a generic stack of data management functions and the need to use synchronization in the common path. These properties are important in particular for fast storage devices, where storage stack overheads dominate and do not allow applications to take advantage of their peek performance.

In this direction DIO will examine techniques for supporting the above aspects:

(a) How can the system achieve protection in the presence of multiple applications that need to share a storage stack and where it is desirable to place part of the data directly in the application address space for flexibility and performance reasons?

(b) What types of mechanisms and operations should DIO provide to support customization of the I/O path for each application (or domain) with limited programmer effort?

DIO will also examine how higher-level abstractions and stronger semantics can be offered for applications without imposing associated overheads to all applications via a generic storage path. DIO will provide libraries for global ordering and synchronization during data access for domains that require such semantics. As discussed above, the core of DIO performs data synchronization within single nodes, by taking advantage of the partitioned address space in key-regions. There are occasions however, where applications may require to synchronize access to data across nodes. Traditionally, such operations are part of the underlying storage system, however, at the expense of paying the associated costs for all applications and common-path operations. DIO will follow the alternative approach of providing such semantics only for applications that need them.

To enable broad applicability of the technology in HPC applications, DIO will also provide a design and implementation of MPI-IO and the HDF5 format (and consequently NetCDF and XIOS that run on top of HDF5). Therefore, we will demonstrate the benefits of DIO's approach both for unmodified and modified applications and we will show how future application design should adapt to take advantage of emerging storage technologies.

APPLICATION QOS AND RESOURCE UTILIZATION

Given DIO's goal to manage fast storage tiers in HPC systems at exascale, it is imperative to provide a resource allocation mechanism. Although some times large-scale systems are used by a single application, in the majority of cases these systems support concurrently lots of applications that launch millions of tasks on different parts of the system but access the same (shared) storage subsystem. Thus, care needs to be taken such that available physical resources are shared fairly among concurrent applications, without sacrificing either application QoS requirements, or hardware utilization. It is particularly difficult to allocate resources properly to such applications, especially in cases where the jobs are created dynamically and their execution time (and other performance requirements) are generally not known in advance. As a result, existing methods take one of two extreme approaches: They either underprovision resources to applications resulting in high resource utilization but poor application QoS or they overprovision which results in high application QoS but poor resource utilization. However, there is need for more sophisticated approaches that will manage to maintain both high application QoS and high resource utilization.

[17] Sudarsun Kannan, Naila Farooqui, Ada Gavrilovska, Karsten Schwan, in Proceedings of 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2014), 23-26 June 2014, Atlanta, GA, USA.

Resource allocation will be even more challenging in the future compared to today due to resource heterogeneity and system size. DIO will design and implement from the ground-up the ability to allocate I/O resources to applications and jobs and to provide the external system allocator and controller that will ensure proper allocation, scheduling, and enforcement of resource boundaries across applications. The main resources that we will consider are: CPU, memory, capacity in each tier, and network throughput. DIO will introduce the following innovations:

- **Monitoring:** The application will only specify the required/projected storage QoS (not amounts of resources) and the system allocator will perform all required accounting (and mapping from QoS to resources) to ensure the corresponding resources are available during application execution.
- **Sharing and isolation:** DIO's scheduler will allocate resources to multiple (and dynamic) applications ensuring infrastructure sharing. However, DIO will ensure via its own mechanisms that (fine-grain) resources allocated to one application are not used by others. This form of isolation is important even in cases where applications access disjoint datasets (as is the common case) because mixed requests to the storage system (for different data) create interference and degrade performance in unpredictable manners. DIO will achieve this isolation via different techniques for each resource, including placement, slicing, and reservation, as appropriate for different resources.
- **Adaptation:** DIO will include a monitoring mechanism to examine if each application is achieving the required level of performance and QoS and to adjust correspondingly the system resources allocated to each job. This is particularly important for applications that dynamically change their behavior over time and not easy to characterize statically. To adapt to changing workloads, DIO is able to move data across nodes and across tiers and also to inquire about data location, if higher layers wish to optimize code (task) placement. Additionally, DIO offers mechanisms to move computation close to data via its APIs. Higher layers can specify what processing needs to happen to data, before data are delivered in application memory by DIO. Therefore, DIO can also take into account "in-transit" processing of data in accelerators.

These mechanisms will allow DIO to both satisfy application QoS requirements but also to maintain high system utilization, as is necessary at the system and data scales envisioned for future applications and services.

IN-TRANSIT ACCELERATION

A common characteristic of a variety of applications that process large datasets is that they typically involve data transformations that despite being simple, they require significant amount of data-movement and processing. For instance, applications may store data in one specific datatype and layout on the device and then require data in memory in another datatype and layout. Another example involves applications, which in order to save space store their data in compressed formats, and they later decompress blocks of interest before deserializing them in memory. All these operations introduce high overheads (that also result in increased power consumption) for two reasons: First they transfer data multiple times between the CPU and memory hierarchies, and second, they keep CPU busy without advancing application state, as the CPU simply transforms data formats.

DIO observes that there is a high potential to eliminate such overheads by offloading common data operations to the storage I/O path that controls all data transfers between memory and persistent storage. The I/O path can be enhanced with accelerator architectures, such as FGPAs, to perform ordinary processing and achieve orders of magnitude higher efficiency for these types of processing. At this point DIO acknowledges that FPGAs are used by only a few Datacenters today[18]. However, given the technology trends, where Intel started shipping Xeon chips with FPGA Accelerators[19], we predict that FPGAs will be part of standard Datacenter equipment in the near future.

However, integrating such computation in the I/O path imposes two challenges: First, the lack of a widely-adopted hardware coprocessing architecture close to the I/O path. Second, the high programmability effort that the co-processing hardware requires and the difficulty to integrate with the existing I/O path. To address these, DIO will provide three important improvements over existing approaches:

(a) An API that allows higher layers and applications to specify computation that should be performed on data during data access. This API can take the form of "co-processors" that are specified once and then used transparently as data are accessed or the form of "tasks" that can be executed on-demand on specific data accesses. Our goal is to explore alternatives and to propose extensions that are both convenient for applications but also allow efficient implementations on future storage systems.

(b) The mechanisms required to communicate data among devices, hosts, accelerators for performing custom computation during data access. These mechanisms are typically the source of high overheads and the topic of current research in several directions. The strength and benefit for DIO is that it controls the data access path and it has the ability to provide efficient support for data-specific acceleration. Although our approach focuses

[18] https://blogs.microsoft.com/ai/2016/10/17/the_moonshot_that_succeeded/

[19] http://algo-logic.com/Intel-Xeon-FPGA

on I/O and does not cover all possible forms and shapes of acceleration, it can significantly improve the efficiency of the I/O path and application performance when data transformations are common.

(c) We will examine and implement specific data transformation functions over FPGA accelerators for the applications in DIO and will evaluate and demonstrate the benefits for future systems.

To achieve this, we will incorporate in our servers and the rack prototype, FPGA accelerators with custom kernels for data processing functions required by the applications in DIO. The FPGA boards have two main advantages. The first one, is that allows the direct connections of NVMe devices to the network bypassing the processors and thus providing lower latency. The second advantage is that they can be used to offload the processor from broad computationally intensive tasks and to provide lower latency and faster responses. In the absence of a mature CPU+FPGA technology, we plan to use for development FPGA board such as Fidus Sidewinder 100 and for demonstration set of lower-cost boards in the rack-scale prototype, such as Alpha Data KU3 that provide the required connectivity and offer high compute power without all the flexibility of the development board.

In addition, we will integrate the use of the accelerators in the I/O path in a transparent manner to the applications. Therefore, our design of the accelerators and kernels will transparently perform concurrent transfer and processing of data by using knowledge of data location and having access to I/O request queues in the DIO storage path.

SKELETON BENCHMARKS: UNDERSTANDING DATA INTENSIVE PROCESSING IN UEABS

Nowadays large-scale real world simulations must leverage parallelism in all different phases. The usual case for leveraging parallelism during the I/O-phase is to use one file per process or one file per MPI task. At exascale this is not ideal because it will lead to hundreds of thousands of files in flight and it can easily oversubscribe the filesystem creating unsurmountable problems with handling massive amounts of file metadata. The alternative to this approach is to perform I/O operations over a small number of processes or MPI tasks. Unfortunately, this will very quickly lead to very poor I/O performance due to synchronization and will significantly downgrade scalability of both I/O and the application.

To better understand these issues and tradeoffs, DIO will build an easy-to-use with minimal overhead profiling tool that will be capable to characterize I/O for future exascale systems. This tool will provide application developers with a representative reflection on how read, write, and metadata operations are executed by the application, offering a more complete understanding of I/O behavior of their application.

Once developed, we will use the profiling tool on The Unified European Applications Benchmark Suite (UEABS). UEABS is a benchmark suite that aims to provide highly realistic and currently relevant codes and data sets that are executed on large Tier-0 and Tier-1 systems. UEABS consists of 12 codes: ALYA, Code_Saturne, CP2K, GADGET, GENE, GPAW, GROMACS, NAMD, NEMO, QCD, Quantum Espresso, SPECFEM3D. These 12 codes cover the following domains: **Particle Physics** (QCD), **Classical Molecular Dynamics** (NAMD, GROMACS), **Quantum Molecular Dynamics** (Quantum Espresso, CP2K and GPAW), **Computational Fluid Dynamics** (Code_Saturne and ALYA), **Earth Sciences** (NEMO and SPECFEM3D), **Plasma Physics** (GENE) and **Astrophysics** (GADGET). The codes were selected to fulfill the same criteria that are required to evaluate realistically technologies that can impact future systems and architectures: UEABS makes available code and suitable datasets, the code does not have any significant barriers to portability and, finally, codes demonstrate good scalability. Equally important, these codes present the majority of the workloads executed by the current HPC installations throughout European HPC datacenters. Therefore, if we randomly pick any European HPC datacenter we are bound to find at least one of these codes as part of the typical workload.

Table 4 shows for all benchmarks apart from Code_Saturne and NEMO that are described in more detail below the input that benchmark uses and whether parallel I/O is implemented using HDF5 or MPI-IO or both. The common workflow for these benchmarks is to create *snapshots* during the execution that are then later used (offline) to do visualization. Using DIO we will be able to couple benchmark execution with visualization process in-situ by manipulating large datasets that do not fit in memory but fit in fast tiers of persistent storage. Using this capability it becomes possible to visualize results of benchmarks in real-time and if they are not satisfactory to abort execution, therefore saving resources (both computation and storage). For example, ALYA uses a lot of I/O during the execution and it is very sensitive to local storage. The amount of data stored for a small-scale mesh (4M) for a snapshot is around 141MB and 196MB. The current state-of-the-art implementation to process this snapshot for visualization using HDF5 takes about 1s for an I/O throughput of 340MB/s. This is well below requirement of 1.2GB/s even for a small-scale mesh to visualize high-resolution 4K images over 24fps.

Table 4: UEABS benchmarks inputs and parallel I/O libraries used by each benchmark. We see that HDF5 and a subset of MPI-IO cover all benchmarks.

Benchmark	Input	HDF5	MPI-IO
ALYA	552.9 million element mesh of generic elements	X	X

CP2K	216 LiH system with Hartree-Fock Exchange		X
GADGET	135 million particles	X	
GENE	Ion-scale turbulence in JET, requiring 3.5-7TB	X	X
GPAW	Calculation of an Au38 cluster surrounded by CH3S ligand	X	
GROMACS	Model of cellulose and lignocellulosic biomass with 3.3M atoms		X
NAMD	3x3x3x3 replication of Satellite Tobacco Mosaic Virus	X	X
QCD	Conjugate gradient solution using Wilson fermions. Lattice 64^3 x 3.	X	X
Quantum Espresso	Functionalised carbon nanotube with a total of 1532 atoms.		X
SPECFEM3D	NCHUNKS=6, NPROC_XI=44 and NEX_XI=1760.	X	X

Furthemore, no significant study has been done on UEABS to characterize its I/O behavior. DIO will characterize in-depth the I/O behavior of UEABS and then will abstract it into a set of smaller and easier to use *skeleton* benchmarks. The idea behind each skeleton benchmark is that it reflects the behavior of the code/application that it represents. In our case the skeleton benchmarks will present all I/O behavior that we have observed in the 12 codes of UEABS. Therefore, when developers write a new application for novel exascale architectures they would be able to run our skeleton benchmarks to evaluate alternatives and identify the best I/O design for their application. Furthermore, when HPC datacenters need to procure a new HPC system they will be able to use our profiling tool and skeleton benchmarks to find out what is the I/O behavior of their top 90% of workloads, map this I/O behavior to one or more skeleton benchmarks, and then during the procurement phase use the selected skeleton benchmarks to evaluate I/O for each procurement vendor. Finally, DIO will use the profiler and skeleton benchmarks to also evaluate the benefits of the proposed technology within the project.

DIO will design skeleton benchmarks based on UEABS that will be used to evaluate the impact of DIO but also will be more broadly useful in characterizing and evaluating current and future I/O technologies.

REAL APPLICATIONS

Besides UEABS and the skeleton benchmarks, DIO will also use real applications from three diverse domains to evaluate its impact: multi-physics, environmental modeling, and agriculture production forecast and analytics.

WORKLOAD I: MULTI-PHYSICS AND MULTI-SCALE ENGINEERING WITH CODE_SATURNE

We have chosen Code_Saturne as one of the project applications because over the past few years it has ran the most out of all other applications on STFC Hartree machines. Production workloads of the Energy Industry in the UK currently depend on it. The energy mix in the country is at the moment: gas 41%, coal 29%, nuclear 18%, hydro and other fuels 5%, wind 4% and oil 1%. Thus, the Advanced Gas-cooled Reactors (AGR) in the UK make up the bulk of the fleet. There is a plant life extension program to continue their usage. In order to ensure safe operation, it is necessary to have a deeper understanding of the flow conditions inside the reactors. The designs are decades old and there is little reliable data from experiments. The objective is to use CFD on HPC systems to get better insight of the flow in fuel assemblies.

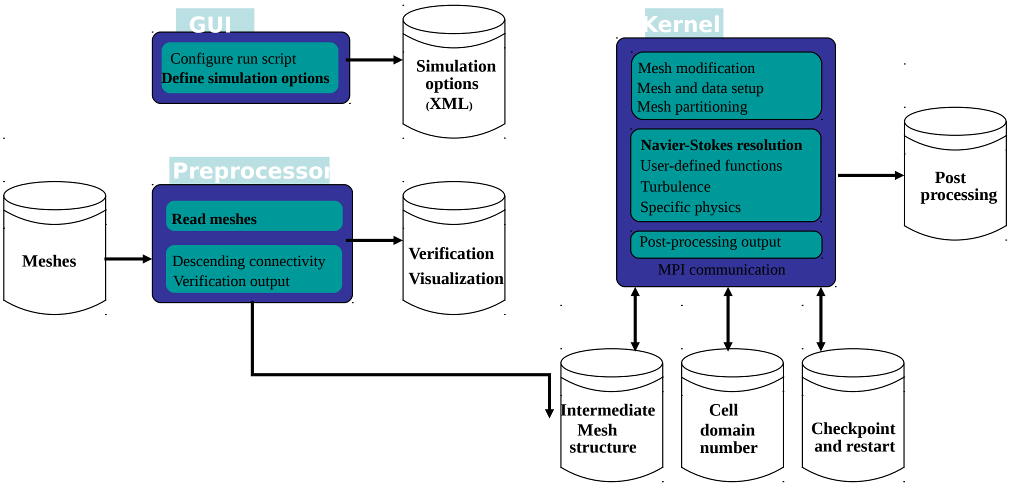


Figure 8: Code_Saturne toolchain and workflow.

The main motivation behind Code_Saturne is that current HPC systems offer possibilities for more detailed configurations in the domains of multi-physics and multi-scale engineering. As a result, a lot of effort has been devoted by the community to optimize solvers (for Navier-Stokes equations) to current petaflop machines. However, fewer resources have been devoted by the application community to explore I/O issues, which is currently recognized as a priority. Figure 8 shows the current workflow and separation of concerns in Code_Saturne. There is a number of tools, where each tool has very rich functionality and there is a natural separation between interactive and potentially long-running parts.

Code_Saturne generates the full mesh for fuel assembly modeling by joining/gluing smaller mesh parts to obtain a 6.5m-high mesh. A periodic pattern exists in the configuration, which is copied and shifted 30 times per fuel element (about 1m high) and a gap mesh is added at the top of the element, on the fly. Then the mesh for one given element is copied 6.5 times, shifted again and glued to get the 6.5m mesh. This mesh has to be dumped to storage, as it is going to be used several times. The smallest 6.5m mesh tested contains about 1B cells, but it is still coarse for this type of simulation. A better mesh would have at least 15B cells to better capture boundary layers. On a system that has 48 racks, where each rack contains 1,024 16-core, 64 bit 1.6GHz A2 PowerPC processors and 8 I/O nodes, which connect to the shared GPFS storage it takes ~730 seconds using 1.5M tasks to generate a 13B cell mesh and dump 2.5TB to disk at a bandwidth of 3.4GB/s. Since simulations can run for very long time (typically days, sometimes weeks), this code also dumps two checkpoint files at a bandwidth of 1.2GB/s.

There are different types of file I/O operations in the Code_Saturne workflow. There are operations to read input, to checkpoint data periodically, to read checkpoints if restarting a previous simulation, and to write output. I/O is so demanding that designers resort to workarounds to reduce impact on application performance and scaling. For instance, input data are not loaded in full during certain stages for: mesh, domain partition (if already known), restart file and input data, while for output is not saved for: mesh (if changed, with added periodicity for instance), domain partition (if computed by the code), listing file, post-processing file, checkpointing and probes. These workarounds do not exist for several phases and they pose restrictions on application evolution.

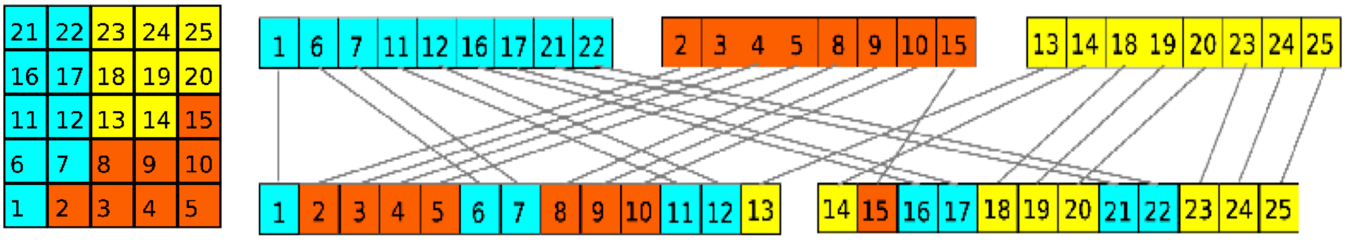


Figure 9: Dataset partitioning (left) and range-based I/O in Code_Saturne, as it will be adapted in DIO.

DIO will examine how Code_Saturne can benefit from the projected technology trends in storage devices, architectures, and the proposed software stack to manage I/O. We plan to examine and use a strategy for I/O in Code_Saturne that decouples I/Os from different processes by using private ranges (over a global namespace), that will then be mapped over DIO’s abstractions both via I/O libraries first and natively subsequently.

For range-based I/O we will use global numbering. Redistribution is on *n* ranges such that the number of ranges is less than the number of cores. Additionally, minimum range size may be set in order to avoid many small ranges and therefore, prevent the storage system from being oversubscribed with files/objects/ranges and metadata. The range-based I/O is illustrated in Figure 9. In this example, the “file” has been divided into 25 ranges. The ranges are then in turn allocated to available cores such that each core gets the group of the ranges with the same color. Once the work has been finished one of the MPI ranks will collect info from the ranges and collate them into the sequential ordering.

DIO has two fundamental advantages in its design that make this approach easy to map: First, it can support large numbers of ranges and concurrent writes. This is inherent due to the way it amortizes write operations from different processes and threads. Then, the index capabilities of DIO (essentially to support scans) automatically sequence the operations without the need for explicit sorting at the application level. Finally, DIO has the ability to reorganize data on the devices and match the different phases of the application.

Our goal is to use DIO technology and enable 10x larger dataset sizes resulting in finer-grain visualization and to couple simulation execution with visualization process in suite in order to quickly assess correctness of simulation to be processed by the Code_Saturne workflow, via both higher I/O throughput and better I/O latencies.

WORKLOAD II: ENVIRONMENTAL MODELING WITH NEMO

NEMO (Figure 10) is a production Earth-climate model application for ocean modeling, developed by the NEMO European Consortium[20] (six EU institutes) and widely used in research and production around the world. It consists of different physical core engines, such as OPA (ocean), LIM (sea-ice), and TOP-PISCES (bio-geochemistry). A 2-way nesting package (AGRIF) and a versatile data assimilation interface (ASM, OBS, TAM) complements the physical engines. Currently, the I/O is managed by the XIOS I/O server. XIOS is a dedicated and parallel asynchronous I/O library, able to deal with the output and diagnostics generation of the model and runs over the NetCDF library using the NetCDF format.

[20] https://www.nemo-ocean.eu

Typical NEMO installations generate vast amounts of data that stress the capabilities of modern storage systems. Currently, there is a requirement and effort to include more numerical models simulating different components of the Earth System, which makes the biogeochemistry model (PISCES) more demanding in computation and I/O. The model uses with thousands of variables in different horizontal and vertical grids, allowing the user to select and output a different set of variables depending on the scientific goal of the experiment. The temporal resolution of these variables also varies, depending on what is being investigated, between hourly to yearly basis (3 orders of magnitude). This combination of number of variables and resolution generates huge amounts of data, overwhelming the capabilities of current systems in terms of I/O throughput and dataset size.

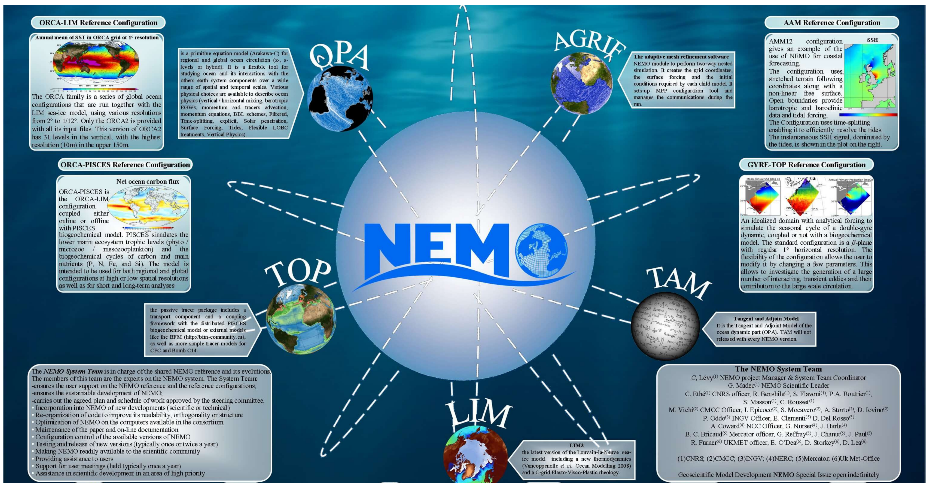

ORCA-LIM Reference Configuration
AAM Reference Configuration
ORCA-PISCES Reference Configuration
GYRE-TOP Reference Configuration
OPA
AGRIF
NEMO
TOP
TAM
LIM
The NEMO System Team

Figure 10: NEMO physical cores. Courtesy of the NEMO System Team.

As an example, CMIP6 experiments output nearly 300 of these fields, distributed among 3-hourly, daily, and monthly data. Modern HPC systems, achieving more than 2 steps per second at this scale, can deliver 3-hourly data every 5 seconds, and daily output, which is of the order of GBs, every 40 seconds. In high-resolution simulations, data increases significantly and puts a lot of pressure on the model that needs hundreds of I/O servers to write data coming from thousands of NEMO compute processors. In today's HPC systems it is not easy to scale I/O for such applications, limiting the type of work and investigations that can be conducted. Table 5 shows indicative problems sizes, dataset sizes, and times for different stages that appear in scientific workflows and where I/O is a bottleneck.

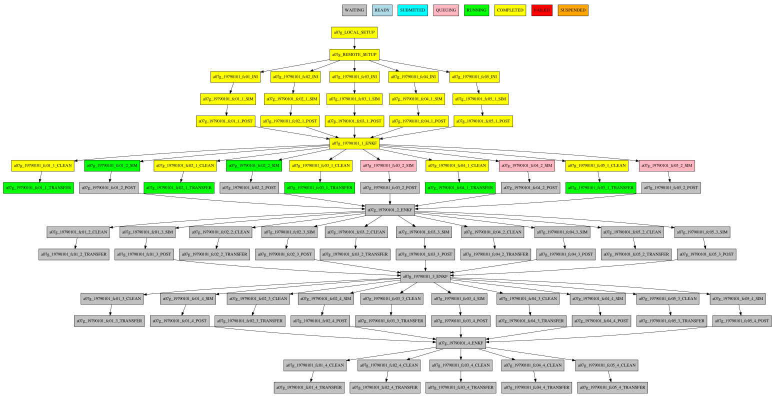


Figure 11: Data assimilation workflow for the NEMO model.

The complexity grows further when designing data assimilation experiments (Figure 11). In this case, for each simulation N members independent with different initial conditions are run, multiplying by N the size of the output. Usually N is a compromise of scientific needs and I/O capabilities.

Therefore, due to limitations in I/O, NEMO saves only a subset of the variables and the data, and thus it limits the reproducibility of its results. For instance, in higher resolutions, we cannot afford to store all ocean levels (usually 75 or 91). Instead, a reduction based in an index is computed and stored. This sampling approach reduces the information delivered to the final user eliminating some of the information that has been computed by the model.

Table 5: Dataset sizes and wall clock time for different stages of a NEMO workflow.

Configuration	Per simulated year data size (GB)	Wall-clock hours per simulated year	Wall-clock hours to visualize raw data	Wall-clock hours to post-process
Standard Resolution (ORCA1)	30	0.9	1.2	0.3
High Resolution (ORCA25)	562	5.3	6.0	1.5

The current NEMO development strategy is to achieve 1/36º horizontal resolution for global simulations by 2022, and simulate a precision of hundreds of meters for regional studies. In the global case this means to increase the biggest configuration available today (1/12 degree) by 27x in computational terms, including the required increase in timestep frequency. This implies to manage a massive amount of data, more than 10x bigger than today.

Data produced in experiments have to be post-processed to fulfill scientific requirements and to be able to produce diagnostics. Diagnostics are really important in modern experiments, for different reasons. In ensembles, where hundreds of members are run, diagnostics are essential to detect which of these members are deviating from the

average or have an unexpected behavior. In tuning and spin-up exercises, diagnostics are key to understand the drift and skill of the model and determine if it is reliable for science. In addition, post-processing can constitute the biggest bottleneck in today's scientific workflows for experiments that produce large amounts of data.
Today's I/O software stack in many cases reduces the opportunity for parallelizing and accelerating I/O and data processing, although workflow and application dependencies allow this. As an example, decadal experiments checkpointing in a yearly basis could post-process all the monthly variables in parallel. Therefore, all processes with a low degree of dependence are suitable to be accelerated using accelerators, FPGAs or GPUs, given the potential they can offer and the ability to overlap data transfer with efficient data transformations.
Therefore, a key observation for DIO is that if we are able to perform on-line post processing and diagnostics that render unnecessary the generation of thousands of GB of raw output, then we will manage to reduce the amount of output generated in production simulations and to move forward with bigger experiments, higher resolution, and higher output frequency. In particular, the use of higher resolution grids is a requirement to solve more processes, improve the interaction between atmosphere and ocean, and resolve baroclinic deformation radius.

NEMO's dependence on the XIOS (XML I/O server) library increases further the requirements of the I/O path because of XIOS's use of the 1-file per process approach for checkpointing. NEMO simulation processes occasionally write checkpoint data (as well as intermediate results), typically 20-100GB twice a day in current setups. NEMO writes each checkpoint in parallel (each process writes its own part of the checkpoint) through a group of XIOS servers (typically on 1-90 XIOS servers for 1000-10,000 core simulations respectively) to netCDF4/HDF5 with compression, to Lustre or GPFS files. To deal with increased failure rates at exascale, these checkpoints need to become significantly more frequent.

DIO will address both of these issues by allowing XIOS, NEMO, and the associated workflows, to operate at 10x larger resolution (e.g. ORCA25 as opposed to ORCA1), taking advantage of larger scale, higher I/O throughput, transparent data transformations in the storage system and the I/O path. DIO will achieve these by modifying XIOS that has control over all data generated and ready by NEMO and the scientific workflows and also allows data transformations to be specified on data that is read or written. In addition, XIOS will benefit from the adaptation of HDF5 in DIO to take advantage of an inherently more scalable checkpointing approach.

WORKLOAD III: AGRICULTURAL PRODUCTION FORECAST WITH CYBELE

Agricultural production forecast is an emerging field that makes extensive use of numerical simulation and machine learning for agricultural modeling and analytics. Today one of the main challenging problems is to automatically identify a crop in a region from satellite images and then to predict crop growth, in order to make accurate in-season forecast of the total agricultural production in one region (Europe, the corn belt, etc.). These two phases of agricultural production forecast constitute the Cybele application and pose significant challenges for the storage system.

Phase 1: Crop identification via processing of satellite images

The application tries to identify the type of crop on soil based on remote sensing data. It is divided in two different stages: the learning stage and the forecast stage.

During the learning stage, the model, based on artificial neural networks (ANN), is trained using time series of satellite images. CYB is a pioneer in this application domain and has worked extensively on supervised learning by comparing outputs with datasets from public agencies but has also experimented with unsupervised learning. Cybele currently uses stochastic gradient descent for learning the ANN. Data from a small region of interest are loaded in memory before the learning starts to issue read I/Os from disk at each computation step of the gradient descent method. This typically limits the quantity of data we can use to ~$5*10^6$ pixels for the learning stage, corresponding to a square of about ~20 km x 20 km. However, it is important to conduct the learning stage on larger regions that are more representative of agricultural production areas. Extending the learning area to a square of about 300 km x 300 km for better matching real needs results in datasets for each gradient computation in excess of ~360 GBytes, which is extremely challenging for today's I/O subsystems.

During the forecast stage, the trained model is used on every pixel of a large region (covering multiple tiles of satellite images). This stage is then schematically represented by the sequence of operations: *read pixel data from disk - compute model from data – write result to disk.* In the forecast stage, currently, the computations are spatially independent (we could imagine trying to improve the detection by looking at spatial correlations) so this step can be run in parallel for every pixel, therefore imposing a large load to the storage system for performing a large number of read and write steps simultaneously. For a single pixel the read size corresponds to two doubles (~20 Bytes) and the write to one integer (code of crop identified) for a total of about 25 Bytes. A typical region of interest has more than 10^9 pixels resulting in more than 25 GBytes of data for answering (query) a single request for crop identification.

Phase 2: Data assimilation in plant growth models

Plant growth models are used to make numerical simulations of plant development in the field during the growing season. The state of the soil-plant-atmosphere system is simulated day after day, taking into account soil composition and climatic data of the day. The state of the system is described by state variables for the plant, such as biomass of the different compartments or leaf area coverage, soil status, for instance with water content in different sublayers and so on. Typically, storing in memory one instance of system state costs ~2 MBytes. Although this is an extremely exciting and important approach that opens up new possibilities, plant growth models have margin errors because:

- Not all mechanistic processes are described in the model; for instance the model will not simulate the effect of pest disasters on crops.
- The mechanistic processes included in the model are never completely accurate because plants are living systems (there is no equivalent of Navier-Stokes equations for plant growth).

To correct for these errors, we use data assimilation methods, much like Kalman filtering for linear systems, using actual data on plant development coming from remote sensing or sensors in the field, acquired during the growing season. Thus, model simulations are corrected online during the season using such data. Since we are not dealing with a simple linear model and we do not particularly expect errors to be Gaussian, we use methods that are numerically more complex than Kalman filters. The method is based on a Bayesian framework where a lot of trial simulations are produced for different sets of model parameters that need to be explored for a better representation of reality. Basically, the method then builds a probability distribution on these parameters and states obtained according to the data measured during the season, repeating for each date of acquisition. The method is called particle filtering because each trial is considered an independent particle.

With this method, large numbers of particles have to be generated to have a correct representation of the actual probability distribution. Studies have shown that for the simplest models with restricted parameter space, the minimal number of particles is of the order of 10^5. It is likely that for our model this will have to be higher, but we have never been able to test that for the following reason: Computationally speaking, it is necessary to store in memory the state of each particle. Given the above-mentioned size in memory of one state, we are rapidly blowing available memory without being able to reach even 1% of the desired number of particles. One solution would be to write to the storage system the states of the particles and process them as sub-sets. Then we could schematically describe the application as a series of operations:

1. Initialize all particles, which requires writings to storage all particle states
2. For the next observation, evolve/update all particle states, which requires reading particle states from storage, compute the simulation then writing new particle states to storage
3. Compute the probability distribution, which requires reading particle states from storage
4. Generate a new particle population according to the probability distribution, which requires writing to storage all new particle states
5. Repeat from step 1 for all observations in the growing season

As part of the Cybele application, we can write to storage particle states in binary format using either ROOT (from CERN) or Apache AVRO. The processes performing I/O (streamers) are part of the library we produce and thus can easily be adapted to using other data formats, for new types of storage, as appropriate. At each step, the amount of data that would be read/written should be a few TBytes to achieve good results. At the same time, the computation itself is highly parallelizable, therefore I/O is currently the main bottleneck for advancing the state of the art. In addition, several instances of this (data assimilation) phase would run concurrently because the data are spatialized (satellite image), with one instance corresponding to one pixel (each run requiring a few TBytes of data and using 1000s of processes/threads).

In DIO we will adapt the models and applications that CYB has developed over time, to use the new storage system and advance the state of the art for crop identification and plant growth modeling for yield forecast significantly. Our goal is to be able to achieve crop identification with models trained in regions of size at least one order of magnitude larger 200x200 km) and to substantially improve the accuracy of plant growth models thanks to Sentinel 1 and 2 data in order to deliver in-season agricultural production forecast with errors lower than 2% on large regions in Europe.

GENDER ANALYSIS

All project partners are aware of the importance of gender issues, especially in technology areas related to computer science and engineering. Over the last few years, individual partners have made efforts and have managed to improve ratios between genders.

In DIO, our goal is to continue these efforts and to further promote work in our area of expertise and related fields towards female researchers and increase their participation to individual groups and for work associated with the project.

1.3 Ambition

DIO aims to innovate and go beyond the state of the art at three different levels:

(a) Overall, to solve an important problem related to storage I/O, moving large datasets between memory and persistent storage (creation, access, manipulation) at unprecedented rates and to enable large-scale, responsive data processing.

(b) At the technology level, to introduce novel systems software technology in the I/O path that will allow storage systems to depart from existing limitations and better take advantage of new device technologies and multi-level storage architectures.

(c) At the demonstration and exploitation level, to provide a working, rack-scale prototype that will show benefits for real applications and to design appropriate Proof-of-Concepts for followup activities.

Next, we discuss each of these levels in more detail.

1.3.1 Fast dataset movement between memory and persistent storage

Although I/O has always been a problem, with the current and projected data growth, many applications are starting to spend a high percentage of time moving data between memory and persistent storage, rather than processing data. New and emerging device technologies offer opportunities to address I/O bottlenecks and maintain an acceptable balance between computation and I/O. The main problem we face in realizing the potential of new storage devices, is that storage system design, including the I/O architecture and the systems software stack require fundamental changes. Merely replacing existing devices with new ones, for instance replacing HDDs with SSDs or introducing additional tiers in the forms of caches is not enough. Such approaches have been the focus of work in the past few years, especially with the emergence of SSDs that, although significantly better compared to HDDs for random I/O operations, they are still relatively slow compared to memory. With the emergence of NVMe devices that further close this gap and the projections towards byte-addressable persistent storage, it has become clear that incremental additions and modifications to current architectures will not suffice. They do not allow us to take advantage of locality and scaling and they impose high systems software overheads, moving the bottleneck form devices to the host CPU.

DIO aims to build a systems software stack that addresses these issues, as follows. DIO allows incorporating storage devices to each compute node. Figure 12 shows our vision on how DIO is going to change future scalable HPC storage architectures, which today follow the paradigm of Figure 1. Today, storage devices, being mostly HDDs and similar size (physical dimensions) SSDs, are placed outside the compute nodes, in one or two tiers. This is mandated by their form factor and the low performance density achieved by a single device (especially HDDs). With emerging NVMe and byte-addressable NVM devices it is possible to place persistent storage, e.g. in the form of small I/O or memory cards, in each compute node. This is made possible due to both the form factor of these devices but also the performance density they achieve. Placing storage devices in each compute node creates potential for more locality during computation and therefore lower latency and also allows I/O throughput to scale with the number of compute nodes.

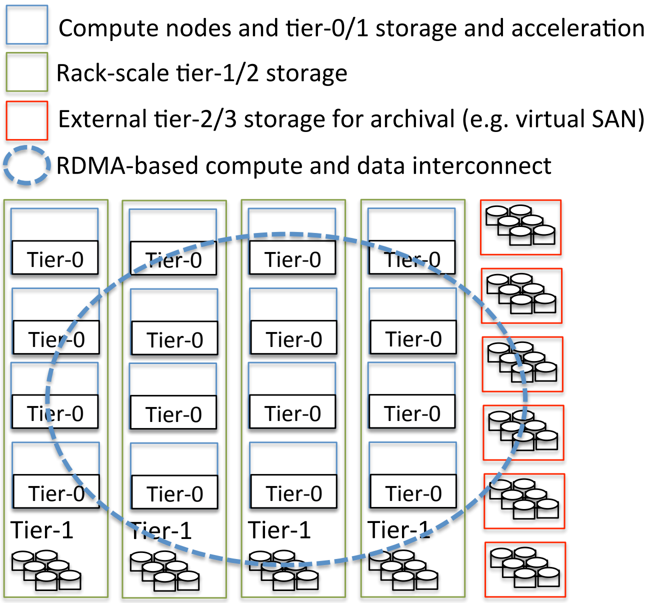

Compute nodes and tier-0/1 storage and acceleration
Rack-scale tier-1/2 storage
External tier-2/3 storage for archival (e.g. virtual SAN)
RDMA-based compute and data interconnect
Tier-0
Tier-1

Figure 12: DIO Enabled Storage Architecture.

However, these two opportunities for locality and scalability are not easy or even possible to exploit with today's storage software stack. It is particularly difficult to introduce a storage tier of fast devices in each compute node and perform replicated writes that tolerate node failures while achieving high performance via the traditional filesystem stack. Existing systems use heavy I/O paths that require going through the OS kernel and using several data management layers even in cases where these are not required by applications. These overheads eliminate benefits from using fast devices, hindering our ability to place local, tier-0 devices within compute nodes.

DIO's software stack inherently incorporates the notion of tiers in its multi-level indexing and data access structures. In addition, the use of memory mapped I/O in the data access path, eliminates the use of the OS kernel and similar cache overheads in the common path. As a result, we foresee that DIO will be able to achieve significant improvements in the performance for moving data between memory and persistent storage. With local tier-0 devices incorporated in the storage system, each compute node can achieve throughput in the order of several GB/s. At a rack scale, this translates to hundreds of GB/s. A large system of 100 racks can achieve I/O throughput in the order of tens of TB/s or about 1 PByte/min, or 1 EB/day. These rates are unprecedented for today's systems that operate with at least an order of magnitude lower sustained I/O throughput.

Besides the ability to quickly move data between memory and storage, DIO will also include mechanisms that will provide deterministic performance for user-facing tasks. This is important in an effort to promote interactive design in existing and new application domains where visualization and inspection of large datasets are required to take decisions and build models.

1.3.2 Introduce new technology

DIO will introduce new technology in the systems software stack for future storage systems. In particular:

DATA ACCESS AND ORGANIZATION

We will provide a new approach to accessing data, based on a new key-value store design aiming and low-overhead and high density by taking advantage of modern device properties.

Although previous work on key-value stores has tried to optimize data access, it has not considered the tradeoff between I/O randomness and CPU overhead. bLSM[21] uses bloom filters to eliminate read amplification and introduces gear scheduling, a progress based compaction scheduler, to limit write latency. FD-tree[22] is an LSM tree for SSDs, which uses fractional cascading[23] to reduce read amplification. VT-tree[24] reduces I/O amplification by merging efficiently sorted segments of non-overlapping levels of the tree. WiscKey[25] improves compaction by keeping values in an external log. LSM-trie[26] uses a hashing technique to replace sorting but does not support range queries. On the contrary, DIO will replace sorting with indexing to reduce overheads and I/O amplification.

NVMKV[27] provides a lightweight, ACID compliant, and high-performance, single-node KV store that cooperates with a host-level flash translation layer (FTL) to store kv-pairs on raw flash. NVMKV uses the lookup structures employed by the FTL itself to locate KV pairs and that are typically hash based, therefore, reducing write amplification. Instead, DIO will use a write-optimized structure and memory-mapped I/O to reduce amplification and overhead without requiring tight coupling with the FTL.

Atlas[28] is a key-value store that aims to improve data-serving density and data replica space efficiency. To achieve these, Atlas employs (1) an LSM–based approach and separate keys from values as to avoid traffic based on values during expensive compactions and (2) erasure coding instead of three-way replication. Atlas targets disk-based storage. DIO shares the same goals for data-serving density and CPU efficiency, but it aims to eliminate expensive compactions, while relying on the use of flash, and to use *m*emory-mapped I/O to reduce overheads.

Previous work of FlashVM[29] and LOBI[30] have focused on the use of flash as swap space for virtual memory. However, this approach is not suitable for persistent key value stores. Applications cannot apply caching policies and pages are not persistent since they are not always backed by flash. SSDAlloc[31] allows the usage of SSD as heap space, and uses its efficient methods to increase sequentiality of writes. SSDalloc could be used for building a persistent key value store, however, an extra protocol should be added to commit SSAlloc's metadata and key value store state. DI-MMAP[32] proposes an alternative FIFO based replacement policy that targets data-intensive

[21] R. Sears and R. Ramakrishnan. blsm: a general pur- pose log structured merge tree. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 217–228. ACM, 2012.

[22] Y. Li, et al. Tree indexing on flash disks. *In the 25th IEEE Intern. Conf. on Data Engineering (ICDE'09)*, pages 1303–1306. IEEE, 2009.

[23] B. Chazelle and L. J. Guibas. Fractional cascading: a data structuring technique. *Algorithmica*, 1(1):133–162, 1986.

[24] P. Shetty, et al. Building workload- independent storage with vt-trees. In *FAST*, pages 17–30, 2013.

[25] L. Lu, T. S. Pillai, A. C. Arpaci-Dusseau, and R. H. Arpaci-Dusseau. Wisckey: Separating keys from values in ssd-conscious storage. In *14th USENIX Conference on File and Storage Technologies (FAST'16)*, pages 133–148, Santa Clara, CA, Feb. 2016.

[26] X. Wu, Y. Xu, Z. Shao, and S. Jiang. Lsm-trie: An lsm-tree-based ultra-large key-value store for small data items. In *2015 USENIX Annual Technical Conference (USENIX ATC 15)*, pages 71–82, Santa Clara, CA, 2015.

[27] Leonardo Marmol, Swaminathan Sundararaman, Nisha Talagala, and Raju Rangaswami. 2015. NVMKV: a scalable, lightweight, FTL-aware key-value store. USENIX ATC '15. Berkeley, CA, USA, 207-219.

[28] C. Lai, et al. Atlas: Baidu's key-value storage system for cloud data. In *MSST*, pages 1–14. IEEE Computer Society, 2015.

[29] M. Saxena and M. M. Swift. Flashvm: Virtual mem- ory management on flash. In *USENIX Annual Technical Conference*, 2010.

[30] S. Ko, S. Jun, Y. Ryu, O. Kwon, and K. Koh. A new linux swap system for flash memory storage devices. In *Computational Sciences and Its Applications, 2008. ICCSA'08. International Conference on*, pages 151–156. IEEE, 2008.

[31] A. Badam et al. SSDAlloc: hybrid SSD/RAM memory management made easy. USENIX NSDI 2011, Berkeley, CA, USA.

[32] Essen et al. Di-mmap–scalable memory-map runtime for out-of-core data-intensive applications. *Cluster Computing*, 18(1): 15–28, 2015.

HPC applications. DIO shares similar goals with the aforementioned systems but has the main novelty to prioritize pages in memory. This gives applications fine grain control similar to user-space application-specific caches.

NextGENIO has proposed a key-value store system[33] for use in Numerical Weather Prediction (NWP) and Climate simulations, as part of the work on next-generation devices. ECMWF[34] is updating their own I/O server software in order to control data access and reduce the number of locks while using fast devices (NVMe). The system they are creating is for internal use, with their collection of tools and for their specific applications, workloads, and datasets, as it is an extension of their current software stack. DIO aims to have a general-purpose I/O stack that can be used to manage fast devices in multi-tier hierarchies (while maintaining the parallel filesystem as the last tier), and for all applications running on the system.

MULTI-LEVEL STORAGE HIERARCHIES

Currently, there are numerous efforts to address deep storage hierarchies and emerging device technologies. Broadly, these approaches can be categorized as:

New storage systems that aim to manage emerging devices and hierarchies. Storage hierarchies in the past have mostly been coupled as different caching layers[35], while only few systems embrace the potential of different storage properties[36,37]. Furthermore, several projects try to integrate NVMe and SCM into the HPC storage stack. For instance, the DAOS project[38] aims to build a scalable distributed object store over fast storage devices. DIO shares the same assumptions about technology trends and the fact that the systems software stack is becoming a main bottleneck. DIO follows the proposed paradigm of DAOS, which advocates the introduction of storage capabilities to compute nodes; however, DIO takes this paradigm a step further and optimizes the storage I/O path of every server. In more detail, DIO will achieve a 2x worst-case performance improvement compared to DAOS as follows:

- One of the main design principles of DAOS, is the backwards compatibility with legacy storage systems, such as POSIX, MPI-IO, and HDF5. Without ignoring legacy formats, on the other hand, DIO focuses mostly on the performance of writes. DIO uses a write-optimized key-value store for data storage and access. This allows DIO to optimize data access and data management operations, to support large numbers of concurrent operations, and to perform fine grain accesses and sharing.
- Unlike DAOS that mitigates concurrency control in the network, DIO decouples and simplifies synchronization by allowing each data item to be managed (at fine grain) by a single storage server and moves synchronization to I/O libraries, and only for applications that require it.
- Unlike DAOS that gets speedup only from the placement of fast storage devices in compute nodes, DIO optimizes further the per-node I/O path in two ways. First by using an optimized memory mapped I/O path with custom (priority-based policies) for optimizing I/O transfers between memory and devices. Second by equipping compute nodes with FPGAs that enable in-transit processing in the I/O path of each server.
- Last, but not least, DIO will be further optimized from the insights that we will collect from the profiling of a wide range of real-life data-intensive applications.

By the end of the project, DIO will achieve a high TRL that has been validated with actual applications.

Application-specific approaches that aim to improve a particular application within existing frameworks. In this category belong several projects[39,40,41,42] that try to use fast devices as burst buffers for transient data (including checkpoints) or that aim to build application-specific storage systems[43,44]. In contrast to these efforts, DIO is building a general-purpose storage system to manage fast devices in multi-tier storage hierarchies.

[33] Simon D. Smart, Tiago Quintino, and Baudouin Raoult. 2017. A Scalable Object Store for Meteorological and Climate Data. In *Proceedings of the Platform for Advanced Scientific Computing Conference* (PASC '17). ACM, New York, NY, USA, Article 13.

[34] https://www.ecmwf.int

[35] Li, C., et al. Nitro: A Capacity-Optimized SSD Cache for Primary Storage. In USENIX ATC, June, 2014. (pp. 501-512)

[36] Wilkes, J., et al. (1996). The HP AutoRAID hierarchical storage system. ACM Trans. on Computer Systems (TOCS), 14(1), 108-136

[37] Yang, Q., & Ren, J. (2011, February). I-CASH: Intelligently coupled array of SSD and HDD. HPCA'2011. pp. 278-289.

[38] Lofstead, J., Jimenez, I., Maltzahn, C., Koziol, Q., Bent, J., & Barton, E. (2016, November). DAOS and friends: a proposal for an exascale storage system. In High Performance Computing, Networking, Storage and Analysis, SC16: International Conference for (pp. 585-596).

[39] NERSC Burst Buffer, http://www.nersc.gov/users/computational-systems/cori/burst-buffer

[40] N. Liu *et al*., "On the role of burst buffers in leadership-class storage systems," *012 IEEE 28th Symposium on Mass Storage Systems and Technologies (MSST)*, San Diego, CA, 2012, pp. 1-11.

[41] Accelerating Science with the NERSC Burst Buffer Early User Program. Wahid Bhimji, et al., In proceedings of Cray User Group (CUG2016), London, Nay 8-12, 2016. https://cug.org/CUG2016

[42] https://glennklockwood.blogspot.gr/2017/03/reviewing-state-of-art-of-burst-buffers.html

[43] Bent, J., Gibson, G., Grider, G., McClelland, B. et al. (2009, November). PLFS: a checkpoint filesystem for parallel applications. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis

[44] Sato, K., et al. (2014, May). A user-level infiniband-based file system and checkpoint strategy for burst buffers. In 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 2014 (pp. 21-30)

Recovery and availability

DIO will make use of copy-on-write recovery to reduce host-level overheads in the data-access path, instead of a Write-Ahead-Log (WAL). WAL produces sequential write I/Os at the expense of doubling the amount of writes (in the log and later in place). CoW performs only the necessary writes; however, it generates a more random I/O pattern. Therefore, although a WAL is more appropriate for HDDs, CoW has more potential for fast devices. The use of CoW is also motivated by three additional reasons; (a) It is amenable to supporting versioning. (b) It allows instantaneous recovery, without the need to redo or undo a log. (c) It helps increase concurrency by avoiding lock synchronization for different versions of each data item, as we discuss in the next subsection.

In terms of availability, DIO departs from previous systems in the use of asynchronous replication, combined with checkpointing semantics. This allows DIO to:

- Reduce the impact of the vulnerability window when a single replica is available.
- Replicate data on a slower tier without impact on application performance.
- Examine the applicability of lightweight techniques and protocols for erasure coding protocols[45,46] that have the potential to reduce the capacity required for replicas on fast and expensive devices.

Checkpointing

Transparent checkpointing of a distributed computation without requiring modifications to user code is possible via approaches such as Distributed Multithreaded Checkpointing (DMTCP)[47] and Berkeley Lab Checkpoint/Restart (BLCR)[48] for LINUX. DMTCP is a widely used technique that does not require modifications to the operating system and supports HPC environments such as MPI, SLURM, and InfiniBand networks. BLCR is a kernel-level system for transparently checkpointing a wide range of parallel applications. *DIO will reduce the size of checkpointed state in transparent checkpointing approaches by using deduplication techniques on the data path. Efficient deduplication is possible in DIO via the content-defined addressing schemes supported by its KV store.*

Previous research pointed out the benefits of a tuned MPI-IO collective approach over the standard "1 POSIX file per processor" approach with the NekCEM application on the IBM Blue Gene/P at Argonne National Laboratory[49]. Other research[50] demonstrated that checkpointing performance in the FLASH I/O benchmark using parallel HDF5 and parallel NetCDF, using an optimized version of ROMIO (MPI-IO) over Blue Gene/L with GPFS achieves high performance, however checkpointing does place a significant load on the parallel file system. Furthermore, checkpointing using parallel file writes via MPI-IO is known to be facing performance tuning issues[51]. Research using ROMIO over the Lustre parallel file system when aggregator processes perform in parallel large, contiguous I/O operations to Lustre shows that write performance (including checkpointing) depends heavily in being properly aligned with the underlying parallel file system. *These results indicate that significant decoupling of I/O from the parallel file system, in line with DIO objectives, is essential in scaling HPC workloads towards the exascale era.*

An existing multi-level checkpointing approach is Scalable Checkpoint/Restart (SCR)[52,53], an open-source library and API that allow applications to take node-level checkpoints in memory, SSD, or HDD within the checkpointing node or at another (partner) node across the network, with parity (XOR), eventually saving checkpoints to the file system to guard against catastrophic full-scale failures. The SCR API is designed to support globally coordinated checkpoints written as a file per process. With SCR an application writes its checkpoint to a system-specified directory (including to local memory or NVM) and notifies the SCR library when complete. Asynchronously, the SCR library manages the data (adds redundancy, etc.) as needed. The Fault-Tolerance Interface (FTI)[54] is another emerging library and API for multi-level checkpointing with similar features. *DIO recognizes the importance of SCR and FTI as emerging interfaces in this space and will deliver implementations of both APIs.*

[45] Heng Zhang, Mingkai Dong, and Haibo Chen. 2016. Efficient and available in-memory KV-store with hybrid erasure coding and replication. In *Proceedings of* FAST'16. USENIX Association, Berkeley, CA, USA, 167-180.

[46] G. J. Akash, et al. 2017. RAPID: A Fast Data Update Protocol in Erasure Coded Storage Systems for Big Data. In *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*(CCGrid '17). IEEE Press, Piscataway, NJ, USA, 890-897.

[47] Jason Ansel, et al., DMTCP: Transparent Checkpointing for Cluster Computations and the Desktop. IPDPS 2009.

[48] Duell, J., et al. (Dec. 2002). *Design and Implementation of Berkeley Lab's Linux Checkpoint/Restart.* Berkeley Lab TR (LBNL-54941).

[49] Fu, J., et al. Parallel I/O Performance for Application-Level Checkpointing on the Blue Gene/P System. IEEE Cluster 2011.

[50] Yu, H., Sahoo, et al. (2006). High Performance File I/O for The Blue Gene/L Supercomputer. *Proc. of HPCA 2006.* Austin, TX, USA.

[51] Dickens, et al. Y-Lib: A User Level Library to Increase the Performance of MPI-IO in a Lustre File System Environment. HPDC'09.

[52] Moody, A., Bronevetsky, et al. Design, Modeling, and Evaluation of a Scalable Multi-level Checkpointing System. *Proc. of the 2010 ACM/IEEE Intern. Conf. for High Performance Computing, Networking, Storage and Analysis (SC '10).* Washington, DC, USA.

[53] Lawrence-Livermore-National-Lab (LLNL). *SCR: Scalable Checkpoint/Restart for MPI.* https://computation.llnl.gov/projects/scalable-checkpoint-restart-for-mpi

[54] Bautista-Gomez, Leonardo, et al. "FTI: high performance fault tolerance interface for hybrid systems." Proceedings of 2011 international conference for high performance computing, networking, storage and analysis. ACM, 2011.

A recent user-level implementation of SCR for MPI named CRUISE[55] stores checkpoint data initially in main memory, transparently spilling it over to other storage, like local flash memory or the parallel file system, using RDMA to pull checkpoint data from compute-node memory into storage servers. While sharing several goals with CRUISE, *DIO features an inherently more scalable design by managing storage levels in an integrated manner through its underlying distributed key-value store. DIO will also build upon and extend recent improvements[56] in MPI collective I/O performance using non-volatile memory for multi-level checkpointing.*

DIO will advance over the state of the art in multi-level checkpointing for HPC applications by drastically reducing host overhead through the use of its mmap interface, its Copy-on-Write (CoW) approach (for versioning, rapid recovery, and low concurrency overheads), and RDMA-based replication. DIO's ability to apply CoW on the update path is one of the key mechanisms for achieving low-overhead, fully asynchronous checkpointing.

DIO will leverage existing standardized APIs, such as MPI-IO, NetCDF, HDF5, XIOS, to achieve interoperability with existing applications, an objective of FETHPC-02-2017 subtopic (c). It will complement existing programming systems such as SCR and FTI by extending them to take advantage of novel converged rack-scale server architectures through integration with the DIO multi-level KV store as well as memory mapped I/O.

DIO will further investigate high-speed replication as a way to achieve increased resilience to failures, and ways to increase the efficiency of replication via use of RDMA and lightweight coding techniques for reducing replica space on devices (avoiding the use of the parallel file system to the extent possible).

Finally, DIO will address challenges in checkpointing state in heterogeneous architectures that include accelerators, aiming for efficient multi-level checkpoint/restart operations for such architectures.

LEGACY AND NEW APIS

Middleware libraries like HDF5 or NetCDF offer a self-describing, hierarchical data organization, which leads to internal structures similar to a file system tree. Data persistence is ensured by interfacing to an underlying file system typically via an I/O library and writing back the file system-like structure to a single file. Unfortunately, the storage layer cannot include semantic information from the middleware and is therefore unable to optimize data accesses based on middleware internal data structures. The resulting typical slow read and write performance can be optimized[57] by delegating nodes within a cluster to act as aggregator nodes, which collect I/O writes from all nodes and which write these I/Os back at a granularity, which is a multiple of the file system stripe size to decrease locking contention in the file system. Alternative approaches like ADIOS reorganize the data layout by using semantic information provided by additional configuration files[58].

DIO uses a different approach, which overcomes the restrictions of block-based file systems. The project uses the internal HDF5 organization (as an example for other middleware libraries) to optimize access to the storage media, moving data being accessed at byte granularity to byte-addressable memory, while keeping bulk data on cheaper flash-based storage or magnetic disks. Even in the absence of byte-addressable storage, DIO naturally overcomes the bottlenecks introduced by today's aggregator nodes by completely distributing data over the cluster, moving responsibility to aggregate data to the node accessing the storage media. The first approach is (partly) related to the DAOS, which tries to directly couple HDF5 with the underlying storage, while our implementation offers a more natural API, which is able to actively include an efficient metadata management based on the key-value API, promising faster speed-ups compared to a standard implementation. DIO, in contrast to DAOS[59], is able to utilize the cluster internal bandwidth, while bypassing today's slow storage stack, promising additional speed-up factors.

DIO will significantly improve the performance of legacy APIs, but will also enable completely new software development approaches. The project will show based on the XIOS-library that simple data annotations based on their access granularity are enough to enable applications to automatically adapt to new storage technologies and to place data within the storage hierarchy exactly where it can optimize a dynamically changeable cost-function. Metadata can also be directly addressed within the key-value paradigm, simplifying and optimizing metadata management, helping to overcome the metadata bottleneck.

IN-TRANSIT ACCELERATION

The advent of FPGA-based flash storage cards enables datacenters and HPC systems to customize their solution for maximum performance, storage capacity, and flash durability. There are currently numerous attempts, both academic and industrial, that use FPGAs to achieve tremendous gains in the I/O path.

[55] Chandrasekar, R. R., et al. A 1 PB/s File System to Checkpoint Three Million MPI Tasks. HPDC 2013. New York, NY, USA.

[56] Giuseppe Congiu, et al.: Improving Collective I/O Performance Using Non-volatile Memory Devices. CLUSTER 2016: 120-129.

[57] M. Howison, et al. Tuning HDF5 for Lustre File Systems, Workshop on Interfaces and Abstractions for Scientific Data Storage, 2010.

[58] J. F. Lofstead, et al. Flexible IO and integration for scientific codes through the adaptable IO system (ADIOS). CLADE Workshop, 2008.

[59] J. F. Lofstead, I. Jimenez, C. Maltzahn, Q. Koziol, J. Bent, E. Barton: DAOS and friends: a proposal for an exascale storage system. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2016.

Hardware acceleration has the potential to achieve higher throughput and lower latency in the I/O path. For example, FPGA-based data-flow architectures[60] and in-transit processing can achieve 80 Gbps/s throughput for certain applications (Memcached, in memory KV store). This is achieved by using a hybrid memory system that combines conventional DRAMs and serial-attached FLASH and optimizing object distribution on different media as well as using customized memory controllers that compensate for large variations in access latencies and bandwidth. BlueCache[61] incorporates FPGA support for handling certain types of requests in in-memory key-value stores and for achieving tighter integration between storage, processing, and memory. Other work[62] has shown how an FPGA platform can be used to accelerate MapReduce applications by offloading the processor from the typical tasks and thus increases throughput and reduces latency, improving significantly the response time of system based on MapReduce by eliminating in most cases the need for access to high-latency storage.

In industry, startup *Attala Systems*[63] has recently presented a novel scheme that uses FPGAs to increase performance in private and public infrastructure (HPC and cloud). Fast NVMe over Fabrics access to flash arrays needs direct access to target drives, bypassing the x86 controller, for lowest latency. Attala Systems has developed a single FPGA, which interfaces between two 40Gbps Ethernet links and M.2/U.2 format SSDs. The FPGA functions as an NVMe target and obviates the need for an x86 motherboard in the target device, lowering cost.

Following these trends, in DIO we will utilize FPGA platforms to allow the direct access of the storage devices to the network (e.g. an FPGA boards with NVMe storage devices and high throughput Ethernet connections – 100 Gbits/s). The utilization of the FPGA boards for the storage devices will allow the hardware acceleration of typical tasks performed by the processor, reducing the response time and improving significantly the total throughput of the storage devices directly to the network. First a design space exploration will be performed in order to find the right architecture based on the HPC applications requirements and then a design methodology will be performed in order to optimize the FPGA architecture that ensures optimum performance of the DIO platform.

Also, FPGAs have a competitive advantage compared with fixed hardware acceleration; reconfiguration. That means that the architecture of the FPGA platform for the DIO can be customized dynamically based on the traffic requirements and the application requirements. These tasks will be performed in WP5.

SHARING, ISOLATION, AND SCHEDULING

The resource schedulers most commonly used in HPC environments, such as PBS[64] or SLURM[65] do not offer support for I/O scheduling and, as such, they cannot enforce basic QoS constraints such as I/O performance and resource isolation, which are indispensable for the controlled sharing and scheduling of I/O resources. Moreover, the resource allocation performed by batch schedulers is generally static, which means that the resource partition for a particular job is selected when the job is submitted and is never changed afterwards.

Due to this fixed resource allocation, jobs that cannot fulfil their requirements with the current resource partition either need to continue running with a non-optimal set of resources or need to be cancelled and resubmitted with an enlarged resource partition. While the former case will probably make the batch job take longer to complete, the latter will lead to scheduling inefficiencies if these augmented resources are only temporarily required by the job. An application's I/O, therefore, becomes unpredictable[66] as it is typically not taken into account by the scheduling process. In addition, the Operating System's I/O scheduler can also cause performance instability and, thus, preclude accurate performance predictions. For instance, two applications writing at the same time to the same storage device (or any other shared resource) will provoke congestion and interference that prevent effective resource utilization. Though this problem has been partially addressed with centralized control and I/O coordination in aggregator nodes and before the parallel filesystem[67], no solution manages to provide a satisfactory view of the whole system and I/O operations, which is one of the goals of DIO.

To tackle these problems, DIO will provide a special scheduler out of the path of the static behavior of SLURM that will be able to understand applications' I/O requirements. These I/O requirements, as well as the batch job's I/O usage and performance, will be continuously monitored and the scheduler will provide the capabilities to dynamically grow or shrink the job's storage resources so that they match with the application's requirements. By relying on this specialized scheduler, DIO's middleware will be able to migrate and transform data to better adapt

[60] M. Blott, et al. Scaling Out to a Single-Node 80Gbps Memcached Server with 40Terabytes of Memory. HotStorage 2015

[61] Shuotao Xu, et al. (2016). Bluecache: a scalable distributed flash-based key-value store. *Proc. VLDB Endow.* 10, 4 (Nov. 2016), 301-312.

[62] C Kachris, et al. An FPGA-based integrated mapreduce accelerator platform,Journal of Signal Processing Systems 87 (3), 357-369.

[63] https://www.attalasystems.com/

[64] www.pbspro.org

[65] slurm.schedmd.com

[66] Dorier, M., et al. CALCioM: Mitigating I/O interference in HPC systems through cross-application coordination. IPDPS *2014. IEEE.*

[67] Gainaru, A., et al.. Scheduling the I/O of HPC applications under congestion. IPDPS 2015. IEEE.

to running batch jobs, therefore decreasing their latency and providing a better resource isolation. This type of support is deemed very important for future schedulers but hardly exists today.

Though profiling and prediction of performance in terms of I/O covered by Dorier[68] *et al.*, their approach is limited to single files and does not take the actual storage architecture into account. DIO will maintain an up-to-date view of the system usage by extracting I/O information from both instrumented and non-instrumented applications. This information will be fed to the middleware and will be processed in order to detect where an application's I/O bottlenecks are and to help deduce appropriate resource allocation to alleviate them. Our prediction approach will use advanced techniques, such as pattern-matching in I/O[69], based on the data captured by DIO.

1.3.3 Demonstration of actual prototype with real applications and Exploitation

Overall, the ambition of DIO is to prototype and deploy its software stack over rack-scale systems, demonstrating unprecedented times for accessing application datasets and then outlining the subsequent Proof-of-Concepts (PoCs) that can be performed in operational environments to further facilitate deployment on real systems. To show the benefits of the proposed approach compared to the current state of the art, we will take a set of actions:

- DIO will build the systems software stack of Figure 4 that will show how the benefits of such modifications and how the future large-scale systems should organize the I/O path.
- DIO will provide APIs and semantics to show how applications can benefit from the new I/O capabilities.
- DIO will demonstrate at the rack level the benefits of the proposed technology and will project to exascale.
- DIO will use real workloads from applications that are popular (Code_Saturne in multi-physics, NEMO in environmental modeling) and forward looking (Cybele in agricultural production forecast and analytics) to demonstrate benefits and facilitate deployment. With respect to applications, our ambition is to achieve and demonstrate significant improvements within the project:
  - Code_Saturne: Finer-grain interactive visualization at 10x larger dataset sizes.
  - NEMO: 10x larger model resolution.
  - Cybele: 10x increase in region size during training and 2% forecast error for prediction.
  - While reducing checkpoint intervals from hours to minutes.
- Finally, DIO will provide plans for follow-up Proof-of-Concepts in operational environments and a technology roadmap with how the proposed I/O systems software stack is expected to benefit operational environments.

[68] Dorier, M., et al. (2014, November). Omnisc'IO: a grammar-based approach to spatial and temporal I/O patterns prediction. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (pp. 623-634). IEEE.

[69] Ramon Nou, et al. Automatic I/O scheduler selection through online workload analysis. In *Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2012 9th International Conference on* (pp. 431-438). IEEE.

2 Impact

DIO will have significant impact in the following directions:

- **At the technical front, DIO will provide innovative solutions to data storage and access for future infrastructures and demanding applications, allowing them to keep-up with data growth and enabling interactive operation.**
- **At the awareness front, DIO will provide better understanding of technology problems, solutions, and limitations for technical and non-technical audiences.**
- **At the exploitation front, DIO will deliver a working prototype with a clear path for its integration to operational environments and for the deployment of the technology in future European flagship HPC systems and by various industries.**

2.1 Expected impact

This section describes how DIO addresses the expected impact of the FETHPC-02-2017 call in general and in particular the subtopic (c) exascale I/O and storage in the presence of multiple tiers of data storage and fast storage devices.

Contribution to the realization of the ETP4HPC Strategic Research Agenda, thus strengthened European research and industrial leadership in HPC technologies.

First, DIO will contribute in several milestones with availability date of 2018 or later of the "Balance Compute I/O and Storage Performance" technical priority of the latest ETP4HPC Strategic Research Agenda (SRA-2). DIO will help meet other milestones of other technical priorities. Table 6 summarizes the milestones of SRA-2 that DIO will contribute, the respective technical priority topics through which DIO will contribute to the milestones, and a short description of the DIO contribution in each case. Next, we discuss first the priority topics to which DIO contributes directly and in a significant manner and then topics to which DIO has secondary contributions.

Table 6: List of Milestones and Topics in the ETP4HPC Strategic Research Agenda, where DIO contributes.

SRA-2 milestones	Research subject	Related technical priority topics	DIO contribution
Direct contribution of DIO			
M-BIO-1	Tightly coupled Storage Class Memory IO systems demo	Topic 5.5.5 Multi-tier storage	DIO provides a software stack to reduce overheads over SCM and fast storage devices.
M-BIO-3	Multi-tiered heterogeneous storage system demo	Topic 5.5.5 Multi-tier storage	DIO provides the software stack to manage multiple tiers and heterogeneous devices.
M-BIO-4	Advanced IO API released: optimized for multi-tier IO and object storage	Topic 5.5.5 Multi-tier storage	DIO builds a system that inherently supports new, higher level APIs, based on key-value store operations appropriate for handling large amounts of data.
M-BIO-6	'Active Storage' capability demonstrated	Topic 5.5.2 Active Stor./On-the-fly/in-transit data manipulation	DIO incorporates in the I/O path the use of accelerators for performing transparent data transformations on datasets
M-BIO-7	I/O Quality of Service capability	Topic 5.5.6 Fine-grained shared Quality of Service	DIO designs scheduler extensions to manage application QoS and resource utilization
M-BIO-8	Extreme scale multi-tier data management tools	Topic 5.5.10 Data Management	DIO designs a software stack to manage multiple tiers of storage
M-BIO-9	Meta-Data + Quality of Service exascale file I/O demo	Topic 5.5.1 Extreme data processing	DIO contributes a working rack-scale prototype and extrapolation to larger systems

M-BIO-10	I/O system resiliency proven for exascale capable systems	Topic 5.5.12 Manageability	DIO provides a checkpointing mechanism tailored for fast devices and multi-tier storage hierarchies as well as a reliable and available I/O software stack
Indirect contribution of DIO			
M-BIO-5	Big Data analytics tools developed for HPC use	Topic 5.5.1 Extreme data processing	DIO with its storage stack, facilitates the development of tools that require switching datasets between memory and persistent storage, e.g. 1) for training and 2) for interactive design.
M-BDUM-MEM-1	Holistic HPC-big data memory models.	Topic 5.1.6 New application domains	DIO provides an I/O path for efficiently moving data between memory and persistent storage in Big Data applications.
M-BDUM-MEM-2	NVM-HPC memory and big data coherence protocols and APIs.	Topic 5.1.6 New application domains	DIO will provide an I/O API to use NVM in big data applications as persistent storage.
M-BDUM-DIFFUSIVE-2	Big Data–HPC large-scale prototype	Topic 5.1.6 New application domains	DIO will facilitate scaling I/O towards a large-scale prototype.
M-ARCH-7	Exascale system energy efficiency goals reached	Topic 5.1.4 Global Energy efficiency	DIO reduces the amount of CPU required for I/O and therefore, leaves more cycles for use by applications. Furthermore, it also minimizes the amount of memory traffic between CPU and storage.
M-SYS-OS-3	New memory management policy and libraries	Topic 5.2.1 Operating System	DIO examines efficient data layout transformations when data move between memory and persistent storage reducing expensive in-memory data movement.
M-SYS-VIS-4	High dimensional data, graphs and other complex data topologies	Topic 5.2.5 Visualization software	DIO provides the storage system that can be used to realize such applications.
M-PROG-API-2	APIs and annotations for legacy codes	Topic 5.3.1 Intelligent Data Placement	DIO provides API extensions related to persistent data placement and management.
M-PROG-RT-5	Scalable scheduling of million-way multi-threading	Topic 5.3.2 Malleability and Dynamic Load Balancing.	DIO takes into account in its scheduling mechanisms and policies, issues related to high degrees of concurrency, sharing and isolation of resources.

A) DIRECT AND SIGNIFICANT CONTRIBUTIONS

TECHNICAL PRIORITY: BALANCE COMPUTE, I/O AND STORAGE PERFORMANCE

Topic 5.5.1: Extreme data processing

DIO will build technology that directly addresses issues and limitations related to extreme data processing. This topic describes the need for unification of traditional HPC systems, such as modeling and simulations, with Big Data Analytics workloads. DIO uses as drivers of the technical work, both traditional HPC applications from environmental modeling that require large datasets and Data Analytics Workloads (Cybele). All applications we use in the proposal already come with large datasets (Section 1). These are **exactly** the types of applications that future HPC storage and I/O technology needs to improve so they can deliver faster results for larger datasets, enabling interactive design. In addition, STFC will provide and use a production level suite of procurement benchmarks in WP5 (UEABS), which include scientific computing applications that increasingly are used as part of Big Data workflows. DIO will use this suite to first understand the impact of I/O in the context of multi-tier storage systems with fast and heterogeneous devices, and second to evaluate the impact of DIO in extreme data processing. We expect that DIO will provide the systems software stack for the I/O path for future extreme data systems.

Topic 5.5.2: Active Storage / On-the-fly / in-transit data manipulation / In Situ processing

According to topic 5.5.2, which extends topic 5.5.1, HPC storage systems need to eliminate data transfers between storage devices and compute nodes due to performance concerns; thus, new storage systems need to be able to process data "in-situ". **In-situ processing is in the core of the DIO** philosophy and it will be achieved in at least two ways. First in WP2, DIO will build a software stack that will optimize the use of multi-tiered storage hierarchies in compute nodes, and thus minimize data transfers over the network. Second, in WP6, FPGA accelerators will enable data transformations, both in terms of layout and content, as data move towards higher storage tiers.

Topic 5.5.3: Energy reduction

DIO addresses energy reduction by 1) increasing density of I/O, and 2) using FPGAs for in transit processing. A main problem today is that storage infrastructure is not used efficiently and therefore, results in high overheads and high energy requirements. One of the main benefits of DIO is that it will increase I/O density, allowing the same infrastructure, e.g. storage devices and CPUs, to serve more I/O requests, significantly reducing "energy to solution". In addition, the in-transit processing of data (layout and content transformations) using FPGAs (WP6), can achieve significant energy savings by eliminating costly (power) data transfers between storage and CPU for ordinary processing. Therefore, we expect that DIO will have a significant impact on overall energy reduction, by achieving higher I/O density and infrastructure utilization.

Topic 5.5.4: Resiliency and Reliability

The topic on resiliency and reliability recognizes the high likelihood of failures during long HPC application runs and the need of mechanisms that will minimize application downtime, speed up recovery, and reduce impact of checkpointing on the underlying parallel file system. The challenge is to reduce the overhead of checkpointing for performance, energy, and rapid recovery reasons. **DIO will contribute a novel multi-level checkpoint mechanism applicable to both legacy and emerging applications** led by partner BULL, an industry leader (WP3, WP6). DIO will address current limitations in two directions: First, the use of fast storage devices in a seamless manner to reduce checkpointing time, allowing higher frequency of checkpointing while ensuring checkpoint integrity, delivering improvements in both recovery-time and recovery-point objectives. Second, DIO will address heterogeneous checkpointing for systems that include accelerators.

Topic 5.5.5: Multi-tier storage

The topic observes the vast diversification of storage devices, and recognizes the challenges of how all these devices can be deployed in HPC. **DIO considers a deep storage device hierarchy** that consists of DRAM, NVM, SSD, and HDD. In fact, DIO will take advantage of the growing availability of fast NVM devices on compute nodes, while transparently managing accesses to those devices. This will allow I/O throughput to scale with the number of compute nodes, an important shortcoming of today's HPC I/O subsystems that use a separate set of I/O nodes. DIO will handle all accesses to the deep storage hierarchy transparently achieving high device utilization and scalable I/O performance.

Topic 5.5.6: Fine-grained shared Quality of Service

The topic recognizes that when millions of processes attempt to access storage concurrently, they will severely impact the performance of the system. Thus, it calls for proper isolation and scheduling techniques that will prioritize tasks when accessing storage concurrently. The core technology of DIO for data storage and access has a fundamental advantage over traditional approaches to file metadata management: it allows DIO to achieve high device utilization even in the presence of large numbers of concurrent I/O requests due to the way metadata and data are organized. This is one of the driving forces behind the use of key-value stores in the Cloud, where there is a large number of concurrent requests from the same or different services. In addition, beyond data storage and access, DIO will design the mechanisms and policies required for a single or multiple applications to allocate resources based on QoS requirements, to ensure that these resources become available when required, and to achieve high resource utilization in the infrastructure by ensuring that resources do not generally remain idle. Finally, DIO takes into account in its design, issues related to workload isolation as part of its allocation and scheduling contributions over SLURM (WP4), so that independent workloads do not interfere with each other when accessing private resources. *Overall, DIO addresses this important topic and problem for future systems at the core of its design and the proposed technology.*

Topic 5.5.7: Layouts/Views/Transformations of data

The topic recognizes the problem of data formatting and the fact that multiple applications need different views or layouts of a dataset, which today leads to unnecessary replication. DIO realizes an important step towards middleware that transforms data on the fly, by providing the ability to perform targeted data and layout transformations in the I/O path that will save significant processing, memory, and interconnect resources compared to when they are performed by the CPU. DIO's APIs and I/O libraries (WP3) include extensions for higher layers to specify data transformations that will be performed in-transit and in fact will be able to also use FPGA acceleration (WP6) and optimized data transfers between storage devices, host memory, and accelerators. I/O libraries on top of the key-value store will be able to specify layout and data transformations and therefore, make efficient use of in-situ processing capabilities. A case in point will be the porting of XIOS on top of the DIO storage system, as XIOS can benefit from layout and data transformations.

Topic 5.5.10: Data Management

This topic recognizes the significance of existing storage systems with evolving data usage features. Although this is a difficult aspect for storage systems, DIO contributes to addressing today's limitations with the use of a key-value store as the main storage engine. The last 15 years, the conventional Big Data community studied extensively the use of key value stores to organize their data[70]. DIO exploits contributions from KV-designs in a way that can evolve as data usage changes. Our key-value store design allows for extending metadata without the need to change how metadata is organized on devices, it allows data layouts to be transformed on devices, and also provides different views on the data to be specified on demand. These are abilities that are not easy to provide in traditional HPC storage systems (e.g. parallel file systems) and can have a significant added value for applications that process large dataset which are required to persist over a long period of time.

Topic 5.5.12: Manageability

The topic recognizes the limited state of the art in *resource management* of HPC data in scale-out deployments. DIO will address this limitation with combining the KV store (WP2) with the resource allocator and scheduler (WP4). DIO's approach focuses on managing data over different device technologies and deep storage hierarchies in a light-weight manner. For instance, DIO provides applications the ability to inherently write a dataset first to the fastest tier and then propagate it to slower tiers based on policies (WP4). In general, DIO provides the mechanisms to move data between tiers (promote and demote) and allocate resources in a dynamic manner and based on QoS. This ability to manage the location of data will allow applications to benefit from large throughput (and low latency) of local devices and the large capacity of slower devices.

B) SECONDARY CONTRIBUTIONS

TECHNICAL PRIORITY: HPC SYSTEM ARCHITECTURE AND COMPONENTS

Topic 5.1.2 Data Access – HW components

Topic 5.1.2 describes the "memory wall" effect, which is the inability of a system to perform in its full potential because it is attached to slow memory. The topic suggests two ways to address the memory wall. First, systems with powerful accelerator hardware can be enhanced with fast memory technologies that can be embedded to accelerators. The second way involves the exploitation of next generation NVRAM technologies whose performance is expected to approach DRAM levels. DIO envisions the use of NVRAM as storage close to DRAM and explores how this can be achieved transparently and at low overhead. Therefore, DIO contributes significantly to our understanding of how memory and fast storage might be unified.

Topic 5.1.4 Global Energy efficiency

DIO contributes to this topic, by addressing issues directly related to "Topic 5.5.3: Energy reduction".

Topic 5.1.6 New application domains

The topic mentions High Performance Data Analytics as the biggest example of a new application domain for HPC. Later in this section we discuss the impact of DIO on HPDA, as part its impact on emerging HPC markets.

TECHNICAL PRIORITY: SYSTEM SOFTWARE AND MANAGEMENT

Topic 5.2.1 Operating System

The topic realizes the complexity of HPC systems that consist of heterogeneous elements and calls for contributions at the Operating System level. One of the dimensions involves memory management of complex memory hierarchies. DIO contributes to this topic by providing a novel systems software stack for the I/O path that is able to manage deep storage hierarchies, including local storage devices to compute nodes and to achieve low

[70] Chang, Fay, et al. "Bigtable: A distributed storage system for structured data." ACM Transactions on Computer Systems (TOCS) 26.2 (2008): 4.

overheads when storing and accessing data, e.g. creating or reading large datasets.

Topic 5.2.4 Resource management and job scheduling

DIO directly contributes to this topic. DIO addresses issues related to concurrent accesses to storage by the same or multiple applications and users. DIO will interface with a popular scheduler, SLURM, and will provide extensions to specify application requirements in terms of storage and I/O and ensure that resources are available for the application before it starts execution. In addition, DIO will monitor I/O performance and will provide mechanisms and policies for taking corrective actions when necessary to either meet application QoS (with respect to I/O) or increase resource utilization (e.g. if it drops below a desired threshold).

Topic 5.2.5 Visualization software

DIO indirectly contributes to this topic by offering to scientists and engineers the ability to load large datasets in memory at extremely high rates, and thus enabling a more real-time mode of application operation.

TECHNICAL PRIORITY: PROGRAMMING ENVIRONMENT

Topic 5.3.1 Intelligent Data Placement

This topic stresses the importance of necessary abstractions that will allow applications to optimize data movement. DIO will contribute directly in this topic in the following ways: First, it will enable the transparent use of local devices to quickly absorb data that is newly created, e.g. by checkpoints or new application datasets. Then, it will manage subsequent data placement to slower tiers for capacity purposes, providing the ability for the application to explicitly move data between tiers. Finally, DIO will provide detailed monitoring of where the data is located and the cost of associated accesses so the system and applications themselves can optimize placement. These are all significant capabilities compared to today's systems that are difficult to extend with knowledge about device types and/or allowing applications to perform explicit data placement on different device types.

Topic 5.3.2 Malleability and Dynamic Load Balancing

DIO indirectly contributes to this topic by addressing resource allocation and scheduling issues in the I/O system.

TECHNICAL PRIORITY: ENERGY AND RESILIENCY

Topic 5.4.4 Use the new levels of memory hierarchy to increase resiliency

DIO indirectly contributes to this topic by addressing the topic "**Topic 5.5.4: Resiliency and Reliability**", as discussed above with checkpointing for multi-tier storage systems.

TECHNICAL PRIORITY: BIG DATA AND HPC USAGE MODELS

Topic 5.6.2. Data Centric Memory Hierarchies/Architectures

The topic realizes that Big Data applications have different memory requirements, as their priority is the minimization of data transfers. The challenge and opportunity for the HPC community is to provide to Big Data runtime systems with data models that they know how to handle and resolve any consistency related conflicts due to parallelism. DIO contributes to this by providing an optimized I/O path for loading data to memory or creating new data on persistent storage. This is an essential component of achieving efficient big data processing.

Successful transition to practical exascale computing for the addressed specific element of the HPC stack

DIO directly addresses this goal by providing a storage stack that is able to manage fast local devices. Typically, today, storage devices (SSDs and hard disks) are placed "outside" compute nodes and applications interact with storage via a set of I/O nodes. However, this cannot scale with the number of nodes at the exascale level. On the other hand, the addition of local storage devices in the form of NVM is projected to allow I/O performance to scale with the number of compute nodes. But these local devices are difficult to use and manage with traditional parallel file systems due to their design and associated overheads. For instance, metadata updates in traditional HPC storage systems involve the use of expensive protocols that do not scale with the number of compute nodes (for this reason today these protocols are limited only to I/O nodes). DIO designs a new software stack for the I/O path that limits the use of the parallel file system to the last (shared) tiers for which it has been designed and manages the additional layers of the storage hierarchy (NVM and FLASH at the node or the rack level) in a transparent manner allowing I/O throughput to scale with the number of compute nodes. In addition, DIO improves latency by introducing low-overhead techniques for accessing local tier-0 devices.

Covering important segments of the broader and/or emerging HPC markets, especially extreme-computing, emerging use modes and extreme-data HPC systems.

There are two ways through which DIO will cover the most important segments of the HPC market:

1) It will capitalize on the trends in HPC storage and *provide extreme storage* to complement the extreme computing segment. It will demonstrate the results through Code_Saturne, a popular extreme computing workload in physics, and through the UEABS procurement benchmark at STFC.
2) It will *broaden the scope of HPC towards data-intensive applications* by providing new I/O capabilities. In this respect, it will demonstrate via the Cybele application in agricultural forecast analytics how emerging applications in High Performance Data Analytics (HPDA) and interactive workflows can benefit from more efficient I/O.

IMPACT ON EXTREME-COMPUTING

DIO will utilize multiple tiers of storage hierarchies (including local devices at compute nodes) and in-situ processing to **complement extreme computing** by reducing data access latencies. Thus, HPC workloads will have access to storage hierarchies that consist of DRAM, NVM, SSDs, and HDDs. DIO will maximize the impact from the use of those hierarchies and will be able to provide low data access latencies. As an example, a case in point is the three project applications, NEMO, Code_Saturne, and Cybele that will all be able to handle datasets that are not feasible today.

Finally, DIO will also measure the impact on applications with extreme computing requirements through the relevant procurement benchmark suites of partner STFC. In the case of extreme computing, we will measure the DIO benefits through the Scalable Science suite, which consists of fully scalable scientific workloads that will run at full scale. This approach brings DIO close to modern operational HPC environments, making easier to assess its impact on extreme computing.

IMPACT ON CLOUD CONVERGENCE AND ADOPTION

The convergence of HPC and Cloud is an area that is expected to have a significant impact both on technology and applications. Convergence between these two worlds is challenging because the have different originating points: serving a single application over a large infrastructure vs serving a mixed workload over a large infrastructure. As we go forward, both HPC and the Cloud need to make steps towards the other direction because user requirements change and also infrastructure efficiency becomes an issue. DIO will have decisive contributions in the following aspects of this convergence:

High latency of data transfers both in HPC and Cloud: Although cloud providers prefer to underutilize their hardware and offer HPC resources to users for exclusive use, as opposed to their common practice of hardware virtualization, users still need to load their data to DRAM either from their local or a remote disk. Thus, despite the recent availability of fast storage devices, such as NVMs and Flash SSDs, Cloud applications today suffer from poor I/O. DIO will provide technology for better taking advantage and exploiting the presence of fast storage devices; it will leverage a storage access hierarchy that consists of DRAM, NVM, SSD, and HDD, in order to reduce data access. Therefore, this aspect of DIO is important for the Cloud as well and the DIO storage system design can contribute to the convergence of the two worlds.

Hardware failures are too common in the Cloud as well[71,72,73]*:* Applications that are deployed in such an infrastructure need to take action in the presence of failures. A few applications could use runtime frameworks that embrace server failures; examples of such frameworks are Apache Spark, and Apache Hadoop, whose execution overheads, however, are intolerable for most HPC workloads. The alternative solution is the use of checkpoints, where the infrastructure keeps snapshots of the state of applications, and applications resume execution from the latest known checkpoint in the presence of failures. However, checkpoints are a non-trivial mechanism, as their frequency in combination with the number of tasks that run concurrently in a computer cluster will produce data large enough to stress the storage system of the cluster. DIO technology will accommodate checkpoints in the context of multi-node setups with local storage (as part of a deep storage hierarchy) and the presence of FPGA accelerators, which is an important step towards reducing the (operational) cost of failures. This is also an important aspect of the convergence between the two worlds.

IMPACT ON HIGH PERFORMANCE DATA ANALYTICS (HPDA)

A new and rapidly growing market segment of HPC is Big Data Analytics, which is also referred as High Performance Data Analytics. According to a recent report[74], the HPDA Market is expected to grow from USD

[71] https://forums.aws.amazon.com/thread.jspa?threadID=28929

[72] https://serverfault.com/questions/326611/whats-the-mean-time-between-failure-for-ec2

[73] https://static.googleusercontent.com/media/research.google.com/en/people/jeff/stanford-295-talk.pdf

[74] High Performance Data Analytics (HPDA) Market by Component (Hardware, Software, and Services), Data Type (Unstructured, Semi-Structured and Structured), Deployment Model (On-Premises and On-Demand), Vertical, and Region-Global forecast to 2021. ResearchandMarkets, Novemeber 2016. https://www.researchandmarkets.com/research/ggnthk/high_performance

25.71 Billion in 2016 to USD 78.26 Billion by 2021, at a Compound Annual Growth Rate (CAGR) of 24.9%. Based on the same report, *"Software is expected to be the fastest growing component in HPDA market" and* "The ability to ingest data at high rates and provide analytics in the real-time to create competitive advantage is the driver promoting the HPDA technology. The software market includes programming tools, middleware, performance optimization tools, cluster management tools, and fabric management." DIO will provide a range of storage related mechanisms that will improve the efficiency of HPDA applications in general. It will demonstrate these with Cybele, a challenging HPDA application, contributing to achieving EU leadership in the emerging domain of agricultural forecast analytics.

A common pattern in Big Data Analytics is the application of Machine Learning techniques on data collections. Machine Learning computations (also used in Cybele) typically involve two stages that usually take place in distributed environments: a) the training and b) the prediction stage. The most common practice during training is the iterative application of usually simple (i.e. linear) operations with techniques such as Gradient Descent, Stochastic Gradient Descent, and the like. This practice typically mitigates the bottleneck from computation to network and storage for very large datasets, because in every iteration, those linear computations require the transfer of large amounts of state from the previous iteration. Moving to the prediction stage the requirements are different, as the input is small but it requires fast lookups on the large dataset.

Given that data transfers dominate in modern Big Data Analytics workloads, DIO's contributions to the efficiency of the I/O path is a great opportunity for HPC to strengthen its position in the HPDA market.

A case in point is the Cybele application (WP5), which faces these challenges both in training and prediction stage. In training, the application uses Stochastic Gradient Descent to classify large collections of satellite images, and since the dataset is large and data transfers dominate, Cybele currently limits its scope into small geographic areas. In prediction stage, Cybele needs to respond to queries and retrieve data from their large collections in real time. **DIO will show through the Cybele application how it will change the field of HPDA.**

Impact on standards bodies and other relevant international research programs and frameworks.

DIO will contribute to standards for I/O libraries, which is the layer that interfaces applications to storage systems and most appropriate for standardization. In particular, NetCDF[75] and HDF5[76] already have standards or user groups that DIO will follow and interact with, especially in relation to extensions to APIs and semantics for supporting fast devices and deep storage hierarchies. Similarly, for the XIOS library, despite the lack of standardization activity there is an active community to which DIO partners will contribute. Finally, the SLURM[77] scheduler, different checkpointing approaches, and each application domain have user groups and activities around the evolution of the respective components to which DIO will contribute.

European excellence in mathematics and algorithms for extreme parallelism and extreme data applications to boost research and innovation in scientific areas such as physics, chemistry, biology, life sciences, materials, climate, geosciences, etc.

DIO indirectly contributes to this expected outcome by allowing applications to consider larger datasets and keep up with data growth, for instance in the case of NEMO (environment) and Code_Saturne (multi-physics).

2.2 Other Substantial Impacts (not mentioned in the work program)

2.2.1 Reinforcement of European Leadership in HPC

DIO will build a new I/O software stack for handling multi-level storage devices and in particular microsecond-level devices that are not possible to use with today's I/O path. DIO will achieve this via a novel approach of changing how the core of the storage I/O path operates, fundamentally addressing shortcomings of incremental modifications to existing storage systems. As such, DIO will contribute directly to European leadership in HPC:

Leadership in systems software for the I/O path:

An IDC'2010 study report[78] mentions: *"Adequate investments in software will be one of the most important determinants of future HPC leadership"*. Another IDC report[79], finds that "In Europe and worldwide, storage will

[75] http://www.unidata.ucar.edu/software/netCDF/conventions.html

[76] https://support.hdfgroup.org/newsletters/bulletin2006.jun28.html

[77] https://slurm.schedmd.com/meetings.html

[78] A Strategic Agenda for European Leadership in Supercomputing: HPC 2020 – IDC Final Report of the HPC Study for the DG Information Society of the European Commission. 2010.

[79] High Performance Computing in the EU: Progress on the Implementation of the European HPC Strategy. Final Report. A study prepared for the European Commission, DG Communications Networks, Content & Technology by IDC, 2015.

remain the fastest-growing segment of the HPC market at least through 2018." Therefore, DIO will directly contribute to leadership in HPC. DIO will design and build innovative systems software for the I/O path and the storage system that departs from the limitations of today's systems and approaches. As such, it will offer European Industry and HPC centers with invaluable tools for building fast and scalable storage systems. Given that systems software is a main limitation for harnessing the benefits of new device technologies, DIO will offer European Industry and HPC Centers a significant advantage over merely using extensions and modifications of today's (high overhead and less scalable) approaches for accessing storage. European Industry will be able to use DIO's systems software for the next generation systems and also as a tool to further explore and understand parameters with respect to I/O for future roadmaps.

Leadership in seamless integration in HPC of new and upcoming storage device technologies:

A main problem today is that we are not able to take advantage of the benefits of fast storage devices, such as variants of NVM, because the systems software dominates the I/O path. For instance, going though the user-kernel interface is costly, two-phase commit protocols for managing metadata do not scale, or running concurrent applications that compete for device access result in random I/Os. These are fundamental issues that require re-design of the systems software in the I/O path, before we can explore and harness the capabilities of new storage devices and deep storage hierarchies for the benefits of applications and services. DIO will offer European Industry the technology and tools for a significant leap in persistent storage and data access performance. Furthermore, by identifying bottlenecks and showing from task T6.4 and task T6.5 what kind of performance improvement and energy saving is possible in codes from UEABS as well as full applications, we believe that DIO will result in immediate uptake of technology by HPC centers, as these workloads are representative of their future needs.

Leadership in new services:

DIO will enable scientists and engineers to move from batch processing approaches for large datasets towards more interactive exploration. In this respect, DIO will both enable new services and applications, but will also start to affect the mentality of users (scientists and engineers) on what will be possible with future infrastructures. This is an important aspect, especially because users today are not able to experience and realize the possibilities that lie ahead. This limitation in turn hinders users from evolving their applications and services, merely because they are concerned that I/O and data access is (and will) not be available to support their new features and capabilities.

Leadership in synergies between HPC architectures and systems software for future roadmaps:

DIO will provide enabling technology to use fast devices in modern and future storage systems. DIO will explore how these devices can be integrated at the right tier (local or remote) and will provide the necessary systems software to seamlessly handle such devices. As a result, DIO will provide insights on important parameters for how future systems should be designed, their potential, and limitations. Today, these insights are of paramount importance for designing well-balanced and realistic systems that can process large datasets. This will allow industry to gauge future directions by quantifying important aspects and taking educated decisions.

2.2.2 Socioeconomic benefits

Market projections indicate, based on an IDC report[80] that "The Global HPC Technology Market Value is expected to grow $45.81 billion by 2024 at an expected CAGR of 6.05% from 2015 to 2024." Based on another IDC report[81] HPC storage revenues are the fastest growing in Europe with CAGR of 7.7% and for 2018 are projected to reach ~900 MEuros. Finally, a report from Intersect360[82] mentions that "*I/O performance is the #1 satisfaction gap in all segments, even non-HPC enterprise*" and that I/O performance is the most important area to improve. Therefore, innovative software for the I/O path to manage deep storage hierarchies has an important role to play. The socioeconomic benefits of DIO, as a technology project, fall in the directions of creating highly-qualified jobs for engineers and scientists and providing visibility for industry to I/O aspects of future infrastructures that will support the data economy.

Training and jobs for highly qualified engineers and scientists in HPC systems software and architectures

DIO will design and build a new storage system, based on new concepts that are able to take advantage of technology trends in storage device technology, networking, and concurrent processing. Therefore, DIO will offer training and expertise that is important for designing future systems and infrastructures. Additionally, we expect

[80] Global High performance computing (HPC) Storage, Networking & Computing Market, Size, Share, Estimate & Forecast, 2015–2024. IDC. May 2015.

[81] High Performance Computing in the EU: Progress on the Implementation of the European HPC Strategy, Final Report, A study prepared for the European Commission DG Communications Networks, Content & Technology by IDC, Report number "SMART 2014/002", 2014.

[82] HPC Adoption in the Enterprise, Addison Snell, Itersect360 Research, October 2014.

that technology from DIO will find its way in industry roadmaps, making it extremely important to also train the people that will be able to realize and deliver on those roadmaps. CYB and BULL are both industrial partners with strong interest in HPC for future infrastructures (BULL) and applications (CYB) and via DIO they will be able to have access to technology and human capital for addressing I/O issues in future systems. In addition, project academic partners have had extensive experience with transferring technology to industry. For instance, FORTH has developed storage technology in two successive iterations of European projects (iteration 1: IOLANES and Streams, Iteration 2: LeanBigData and CumuloNimbo) that was eventually licensed to industry and is currently part of products in the market. We foresee that DIO will follow the same path, since the technology that will be developed addresses fundamental limitations of today's approaches for data access.

Visibility for Industry in future opportunities and challenges and HPC - Cloud convergence

A 2012 IDC report[83] for market analysis of 2011 worldwide big data technology and services revenue share by segment mentions that Storage amounts to 11.8% of revenue and Software 29.7% (in addition, Servers are 14%, Networking 3.1%, and Services 41.5%). This shows that storage and systems software play an important role in the Cloud market and that new technologies that improve the utilization of storage devices and enable the use of faster device technologies have an important role. However, Industry today is facing several challenges with respect to harnessing the benefits of new device technologies and addressing the limitations of current systems, including keeping up with data growth. DIO will provide technology that will allow industry to understand tradeoffs, explore new possibilities, and eventually design our next generation HPC and Cloud infrastructures and systems software.

2.2.3 Environmental benefits

Although the main goal of DIO is to design new technology, one of the areas where we envision this technology will have a direct impact is environmental modeling and environmental risk analysis applications and services. In particular, DIO will enable:

- handling large datasets,
- in a more interactive manner than what is possible today, and
- performing custom processing efficiently in the I/O path.

These are fundamental capabilities that can allow environmental applications to move from today's datasets and applications towards faster and reliable weather and climate services, allowing to the final user an improved exploitation of the results. As long as the size of the output generate grows, is crucial to maintain or even reduce the time going from the generation of the data to final analysis. World intercomparison projects, such as CMIP6[84], will benefit from DIO improvements, allowing faster generation and processing of climate data. In the same way, developments in XIOS will be reported to the NEMO consortium, allowing all the model users and institutions to take advantage of the enhancements developed. We expect that DIO will significantly enhance our abilities for modeling the environment, understanding risks and challenges, and also enabling new services and applications.

2.2.4 Industrial Impact

DIO will have specific, significant, and continuing industrial impact in at least two fields: IT infrastructure and agricultural forecast:

At the infrastructure level, DIO will lead to enhancements in architectures and systems designed by BULL and operated by HPC centers, such as STFC and BSC, with appropriate storage hierarchies, device technologies, and a software stack, highly tuned for each configuration. Therefore, DIO represents for BULL a potential differentiating factor in comparison to commercial solutions.

As explained in the methodology section, the demonstration application on agricultural production forecast is computationally limited by memory constraints. With improved management of I/O, DIO will enable working on much larger datasets and ideally reaching regional scales. Thanks to the project, the application will then be able to produce agricultural production forecasts during the growing season on large regions in Europe or other parts of the world (Corn belt, South America, Australia, Russia, Black Sea, etc.). The targeted accuracy is of about 2% per region for a forecast between start of grain filling up to 2 weeks before harvest (typically end of August for corn or mid-June for cereals). Mastering such information is crucial for all industrial or public partners of the agri-business. First of all for farmers' cooperatives: they will be able to plan in advance their logistics and put on the market their production at the best prices by anticipating commodity prices. Insurance companies will be able to identify in

[83] Worldwide Big Data Technology and Services 2012–2015 Forecast. Market Analysis by IDC. March 2012.

[84] Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937-1958, doi:10.5194/gmd-9-1937-2016, 2016.

advance when they need to set aside sufficient funds for compensating farmers in the case of a bad harvest. Trading companies will also benefit from anticipation of commodity prices. More generally, the anticipated knowledge of agricultural productions will help decrease volatility of commodity prices and reduce speculation in the market. This in turn will benefit producers (farmers) who will know in advance their income for planning investments, and consumers, especially in developing countries who need to secure their food supply.

2.3 Barriers to maximize impact

DIO aims to address the fundamental problems with I/O and data access in future infrastructures. Besides the technology that DIO will develop and which is at the heart of the proposal, there are a number of other barriers to addressing these problems:

2.3.1 Workforce considerations

Today, there is generally a lack of expertise on designing storage and I/O subsystems that are able to keep up with data growth. In addition, there is a lack of expertise in configuring systems to efficiently operate in large and complex setups. For instance, most of the filesystems used today extensively in HPC infrastructures, including GPFS and Lustre, have been notoriously hard to extend, configure, and tune in different environments, as other components of HPC systems and infrastructures evolve over time.

DIO plans to address such considerations related to the availability of expertise and human capital, by designing technology that is able to monitor system and application operation and to adjust to different parameters and conditions automatically. WP4 will design mechanisms and policies that allow our prototype to adjust to different workloads and conditions. In addition, DIO will also contribute, via WP7 and dissemination activities in educating, scientists and engineers, on principles of future storage systems and what this implies for applications and services. Via self-adaptive technology and dissemination, DIO will address workforce considerations in the area of data storage and access.

2.3.2 Cooperation of other links in the value chain

Another important barrier is that storage and data access are only one part of the value chain in data processing and HPC. At the same time as storage is evolving due to new opportunities and requirements, memory, processing, networking, systems software, and applications software are also in flux and under significant change. In fact, the interactions between these different aspects of modern infrastructures are not easy to handle, as they involve different communities and to some extend industries. To address this barrier, DIO includes partners that are experts at most levels of the HPC software stack (FORTH, BSC, STFC, JGU, CYB), HPC centers that can deploy the foreseen technology in real systems and applications (BSC, STFC), partners at the hardware, both component and system level (BULL, ICCS), and industrial partners that have vested interest in commercial HPC applications (CYB). Through the networks of these partners, DIO can achieve a considerable cooperation among all important links in the value chain to gather feedback from each sector/community but also to provide guidelines that will be useful in their individual goals. For this reason, WP7 includes actions that bring these links and actors together to ensure cross-fertilization in the end-to-end value chain.

2.3.3 Standards and legacy software

Users of legacy software can be challenged to embrace DIO, especially when the APIs of their data are not supported. For this reason, DIO will provide support in its stack (and at different levels) for existing formats HDF5, NetCDF, MPI-IO, and XIOS, which are used extensively in HPC applications, via the corresponding I/O libraries.

The layer of the I/O library is also where most of the standardization activities take place with respect to I/O. NetCDF and HDF5 involve standardization activities, where DIO partners will participate and contribute. XIOS is a more recent effort, however, with an active community (especially for the weather and climate modeling community), where also DIO partners will participate in discussions for future extensions. Finally, the SLURM scheduler and each application domain have user groups and activities around the evolution of these components to which DIO will contribute.

Overall, the planned actions in WP7 are adequate to lead to adoption of the technology from real applications and services. In addition, and after DIO technology has been further validated we expect that new research will expand the scope and applicability of the proposed technology both in HPC and the Cloud. In particular, this type of work typically falls within the work conducted by software teams in HPC centers, domain-specific tools and applications such as environmental modeling, and industry vendors that design, integrate, and support HPC systems. DIO includes partners from all these communities and via the dissemination actions in WP7, DIO will explore the pathways for continuing technology development in these communities.

2.4 Measures to maximize impact

To achieve its goals, DIO foresees dissemination, communication, and exploitation activities with specific targets and towards different audiences. These activities (WP7) aim to maximize project impact as follows.

2.4.1 Dissemination of project results

Dissemination of project activities and results aim to cover the following directions:

Research: To ensure that infrastructure researchers and vendors become aware of the project technology in the area of data storage and access. This will be achieved via research publications to relevant and visible venues, e.g. USENIX FAST, IEEE MSST, HiPEAC, USENIX ATC, Supercomputing, ACM Transactions on Storage, Nature Geoscience, Geoscientific Model Development, that are followed by researchers and industry and have the potential for high impact. In addition, most of these venues include presentations, which are an important aspect of dissemination, and also discussions with participants both from research and industry. Furthermore, we will use research meetings and the individual research networks of project partners to discuss project technology and results as well as to solicit technical feedback.

Technology visibility: To ensure that vendors and policy makers become aware of project technology and how it can affect future system architecture and roadmaps, we will use the collaboration networks of partners and events targeting discussions between related stakeholders. Several partners have open channels of discussion with industry about future trends and technology limitations and will use these to discuss project activities and results and to also solicit feedback.

Domain-specific applications: To ensure that scientists and engineers in different application domains become aware of DIO technology, we will participate with technology presentations in domain-specific meetings, e.g. for applications in environmental modeling, HPC I/O, as well as tutorials on project technologies to offer a deeper understanding of differences and capabilities with respect to traditional approaches. The targeted meetings include, for instance, ISC High Performance[85], NEMO Enlarged Developers Committee[86], EGU[87], AGU[88], SuperComputing[89], and PRACE days[90].

HPC-Cloud convergence: To engage with researchers in Cloud with respect to emerging data storage and access technologies and how they can accelerate convergence of HPC and Cloud, we will organize participation in conferences related to the Cloud, e.g. ACM SoCC[91] and IEEE International Conference on Cloud Computing[92], via publications, tutorials, or workshops with the specific goal to discuss data access convergence between HPC and the Cloud. We will also explore synergies with projects in the area of Cloud (included in Table 2), to discuss requirements and trends and identify the major aspects that can be unified across respective technologies.

Training and education: To provide material for training and education with respect to fundamental issues in data storage and access and the problems that DIO solves, research and academic partners will introduce material related to project technologies in existing and new undergraduate courses, graduate courses and theses, and conference tutorials. The purpose of introducing material at different levels is to raise awareness of requirements and baseline approaches (undergraduate level), technology limitations (at graduate level), and potential solutions (tutorials in two flavors: research-oriented and industry-oriented).

Public awareness: To provide high-level material for the general public on what is the "storage and data access" problem we face today, why and how this can be addressed, and the impact it can have on new applications and services with increasing data sizes, we will use online media, including short web presentations, Twitter and LinkedIn updates, short newspaper articles in specialized and general press, and high-level presentations of technology in YouTube and the project web page.

An important dimension in this direction is Open Research Days that are hosted by project partners (FORTH, BSC, JGU, ICCS, STFC). During these days that happen a few times every year thousands of individuals visit partner premises and are introduced to science and technology. We will use project material from DIO to discuss problems and trends in data storage and access, an area that people tend to find exciting and interesting because most have already experienced problems as individual users of technology. In addition, partners (FORTH, BSC, JGU, ICCS) regularly host school visits that are more targeted than open research days to high school students that may plan or are interested in a career in science and technology. During these visits, hundreds of high-school students have

[85] http://isc-hpc.com
[86] https://www.nemo-ocean.eu/event/2017-enlarged-developers-committee/
[87] https://www.egu.eu/
[88] https://sites.agu.org/
[89] http://www.supercomp.org
[90] http://www.prace-ri.eu/pracedays17
[91] https://acmsocc.github.io/2017/index.html
[92] http://conferences.computer.org/cloud/2018/

guided tours in labs and demos. Our goal is to include DIO material and presentations at a level that will excite imagination and interest.

Standards: To interact with standards organizations and in particular groups that discuss APIs and semantics (HDF5, NetCDF, XIOS, SLURM) to see how DIO technology can seamlessly be used as broadly as possible. Project partners are already involved with such activities, e.g. JGU has participated to Lustre technical meetings and has contributed via its work to improving metadata aspects in Lustre for the purpose of concurrent application access and regulating individual application throughput. DIO will participate to pertinent meetings (as a first step, meetings associated with codes and applications in the project, including ISC High Performance[93], NEMO Enlarged Developers Committee[94], EGU[95], and PRACE days[96] and then extend our reach to additional application domains) to explore how project technologies can develop synergies with existing and upcoming application-domain standards.

EU-wide-synergies: DIO will pursue synergies with other EU projects in neighboring technologies, e.g. programming models, system tools, applications and models, interconnects, system architecture, cloud infrastructures to facilitate technology awareness and visibility but also to pinpoint future problems and challenges. In addition DIO will participate in ETP4HPC and contribute to working groups. DIO members are already active in ETP4HPC (FORTH, BSC, BULL, CYB, STFC, JGU) and through DIO they will extend and increase their level of engagement and they will help broaden the reach and impact of ETP4HPC.

Industry Day[97]: Towards the end of the project, DIO will organize an Industry Day at STFC, using an established practice of STFC based on the InCEPT process that STFC has developed to bring technology and applications together. This event will focus on project technology and how it can address modern and future needs.

In building the DIO community, we will use the collaboration and research networks of individual partners, which will ensure EU-wide coverage. Consortium partners have a strong reputation in their respective areas of activity which ensures broad access to the research community in Europe as well as to European and global industry. In addition, each partner will be responsible for general public and education activities in their geographic area, ensuring breadth. To simplify the dissemination and communication process and provide a uniform graphical profile to the target audiences, we will develop stationary material (leaflets, templates, poster) for public and restricted events (e.g. conferences, workshops, exhibitions, discussions with industry, etc.).

Table 7 shows a list of dissemination actions for each category. Overall, DIO dissemination activities cover the scope of actions required to achieve high impact. In addition project partners have the required expertise and appropriate profiles to carry out these activities in a timely manner, while taking corrective actions where required.

Table 7: Summary of dissemination activities.

Category	Activities Description
Across all	Prepare communication material, including Website, logo, templates, audio-visual material, social media tools (LinkedIn, twitter), mailing list, leaflets, posters (three versions: start, middle, results), flyers (three versions: start, middle, results), newsletter (bi-annual) with project updates.
Across all	Mapping DIO stakeholders and communities and maintaining throughout the project
Research	Presence with publications and presentations to research venues (conferences, journals, workshops)
Technology visibility	• Conduct demos and proof-of-concepts with technology • Contribute to roadmaps with projections related to project technology • White papers • Use the general communication material
Public awareness	• Press releases • Articles in general press • School visits, science fairs, open days • Use the general communication material

[93] http://isc-hpc.com

[94] https://www.nemo-ocean.eu/event/2017-enlarged-developers-committee/

[95] https://www.egu.eu/

[96] http://www.prace-ri.eu/pracedays17

[97] http://www.stfc.ac.uk/files/ensuring-high-performance-research-meets-industrial-needs/

Applications	• Presentations in user-groups and application-specific events
Training and Education	• Tutorials, courses, theses • School visits, science fairs, open days
Standards	• Participate in meetings related to API, semantics, and standardization of I/O libraries and to related aspects of project applications
HPC-Cloud convergence	• Publications is technical conferences for technology convergence in storage and I/O • Discussion with industry for proof-of-concepts with applications from the Cloud
Standards	• I/O libraries (NetCDF, HDF5, XIOS) • SLURM and application-specific groups, depending on potential
EU-wide synergies	• Existing EU projects (as mentioned in Table 2). • Participation and contributions to ETP4HPC and working groups. • Providing speakers, booths, papers etc. at specific events (conferences, training, and workshops) forming the agenda and future of storage and I/O systems. • Participation in panel discussions, meetings, working groups, consultations to contribution to EU storage and I/O vision. • Position papers based on the knowledge and experience from DIO.
Industry Day[97]	• Bring DIO, industry, and applications together • Present technology and discuss potential

OPEN ACCESS SCIENTIFIC PUBLISHING

DIO will fully embrace the H2020 requirement for Open Access publishing, following the guidelines presented by the European Commission. The requirement may limit the venues that can be used for publication but since several venues are now providing Open Access options we do not foresee any problems. We expect that DIO will use both 'green' and 'gold' publishing, i.e. self-archiving in one or more repositories and to some extent through paying an author's fee when publishing.

Publishers provide various options for publication: IEEE, ACM and Elsevier provide several models for Open Access publishing, with various options on what license to place on the published documents. Associated with the 'gold' publishing is a modest cost for the author (and in turn, the funding body), in the case of IEEE this fee varies from $1350 to $1750 per paper depending on which journal is targeted. Publishing fees have been estimated and taken into account in the DIO budget. However, several top venues, e.g. the top-tier Usenix series of conferences, already provide free Open Access for research purposes. Additionally, technical reports and other forms of technical publications can be provided via Open Access through a 'green' approach using e.g. arXiv.org or other appropriate repositories. Similarly, (public) deliverables produced by the project can also be archived in widely used repositories, to ensure the long-term availability of the material even after the end of the project.

The software produced by DIO will be provided to the community *for research purposes* under an appropriate license. The license to use, out of the available options, is not clear at the moment, because the new software stack will include at points dependencies with existing systems and components, and will also depend on other IPR questions concerning specific pieces of software and potential commercial exploitation. These decisions have to be delayed to later stages of the project and around the time components start to become available.

Special care will be taken to ensure that long-term stable repositories are used to provide access to the software in addition to providing a useful interface for other authors to re-use and contribute back to the developed software. There are several popular options that provide long-term availability and easy collaboration, such as GitHub, sourceforge, etc. These resources may be used for both internal development and long-term public access. However, for security reasons, maintaining local repositories for development and later publishing to e.g. GitHub, may be more appropriate, so that any IPR issue can be solved before publishing. Project partners are experienced with several options and will take appropriate decisions at project start.

These issues, i.e. Open Access publishing and software distribution, will be handled in the Consortium Agreement, to ensure that all partners agree in a legally binding manner.

2.4.2 Exploitation of project results

Joint exploitation between academic and industrial project partners

As a whole, DIO will produce a new software stack for data storage and access that can be deployed on future infrastructures to manage the I/O path. In addition, certain components of the software might have additional uses and particular interest for certain sub-sectors of the Industry and certain domain-specific applications, e.g. the FPGA acceleration aspects for applications that use XIOS. Based on the extensive experience of project partners, DIO foresees the following forms of exploitation, listed in increasing risk:

Exploitation in "Proof of Concepts" (PoCs) by supercomputing and industrial partners: Industrial and HPC center partners (BULL, CYB, BSC, STFC) will create plans for 3-4 "Proof of Concept" (PoCs) that will further validate benefits in setups closer to operational environments. DIO's approach has the advantage that it can allow racks of existing equipment to be augmented with next-generation storage devices and accelerators at moderate cost and to create small/medium scale systems of one to a few racks running real applications to validate system benefits in specific contexts. Project partners with visibility and direct interest for application performance in different domains (BULL, CYB, BSC, STFC) will pursue such PoCs, further exploiting DIO technology.

Exploitation in product roadmaps and the Extreme Scale Demonstrator project: We expect that DIO technology will affect the product roadmaps of Industrial partners. Both BULL and CYB are closely associated with commercial HPC products and roadmaps are an important instrument to interact with potential users as well as to validate the feasibility of future plans. DIO, by addressing specific problems that have been tantalizing the I/O path as new storage devices emerge, it will provide a clearer view of what will be possible in the future but also any limitations that will emerge. As such, it will provide invaluable input to product roadmaps in terms of performance and scalability that can be achieved, taking also into account other system parameters, especially storage device evolution. Additionally, this visibility into projected overheads and scalability will provide better understanding of the next set of bottlenecks and the technologies that can be used to tackle these as well. This is particularly useful for product roadmaps today, where uncertainties about I/O overheads obscure next-generation system design. Given these benefits and their potential, DIO will work with Extreme Scale Demonstrator projects[98] to plan the path for deployment of DIO technology.

Exploitation via licensing of technology to industry to create new products: Project academic and HPC center partners (FORTH, BSC, STFC, ICCS, JGU) have previously licensed research technology to industry, beyond the circle of project partners, to create new products. This model of exploitation is more open-ended because it allows inclusion of new technology to a broader range of products. Although it involves extensive legal and financial discussions and is a multi-faceted endeavour, it is an important vehicle for exploiting technology and achieving impact. Given that project partners have experience with the process and they have extensive networks of collaborators in relevant industry, and given the projected importance of DIO technology, we are optimistic that during the execution of DIO similar discussions and opportunities will arise.

Overall, project partners, including academic, supercomputing, and industrial partners, have a vested interest in the technology to be generated because their current products and services are directly related to data storage and access (and processing). Therefore, we believe that there is a lot of opportunity to form joint ventures for further exploiting DIO technology in certain domains that are data and I/O intensive. This approach of joint ventures is particularly promising today and for technology-intensive areas, where a single actor (partner) typically does not have all expertise to pursue a complete, end-to-end solution in a timely manner.

Individual exploitation plans

Table 8 summarizes previous efforts and plans of individual partners with respect to exploitation of research results in DIO.

Table 8: Partner-specific exploitation plan.

Partner	Dissemination plans
FORTH	Having a prestigious track record of successful research and development projects, FORTH will further enhance its position in cutting edge research projects in the area of accelerator architectures and scalable storage systems, two important areas in IT infrastructures and future datacenters. Over the last six years FORTH-ICS has filed for several patents in the area of storage subsystems and interconnect architectures, two important aspects of modern systems. Additionally, CARV (the lab of FORTH-ICS that participates in DIO) has licensed storage technology to industry in two different

[98] http://www.etp4hpc.eu/pujades/files/EsD%20Concept%20-%20Current%20State%2022%20Jul%202016%20-%204.pdf

	rounds, during the periods 2011-12 and 2014-16, as an exploitation step of EU-funded research. Furthermore, in two cases, FORTH-ICS and CARV have attracted development centers of high-tech startups (iofabric.com and kaleao.com) in the Science and Technology Park–Crete (STEP-C) in the areas of datacenter storage and datacenter server design. FORTH will use DIO to identify opportunities that may result into new products and services, as has happened with recent storage and server related technologies.
BSC	BSC will exploit the results obtained in several ways: all developed software will be open source enabling a global exploitation of the results obtained in this project. In particular, changes in XIOS, if they provide a breakthrough on performance as expected, they can be used in production in BSC infrastructures. About the modifications of SLURM and the storage scheduler they will be deployed in the BSC infrastructures used for research. Finally, BSC will promote them for use in other research projects and in contracts with companies.
STFC	STFC will use the DIO project result to feed its processes when conducting detailed analysis of business opportunities within existing and emerging markets. The Hartree Center's focus on Industrial engagement and its already strong links with international industrial and technology partners places it in strong position to inform the evolution and adoption of exascale solutions by providing access to systems, tools and know-how in the pursuit of solving real world problems. DIO will facilitate an extension of the architectural characterisation and evaluation of novel technical solutions that the center is undertaking as part of its future technologies focus and adds value to our industrial engagement across HPC and Big Data. In addition, as the leader of WP5, STFC will work on design and implementation of an easy-to-use tool that will be able to profile in-depth I/O bottlenecks of the future exascale applications. We will be applying this tool on the workloads that our industrial and technology partners are running on HPC systems to improve their performance and energy use. The high-level simulator that will be developed as part of WP6 will be used by our research engineers to demonstrate to customers what impact future technology will have on their codes and to inform vendors during procurement what should be the best-fit architectures for the codes that are about to be run on new machines.
JGU	Big data applications, huge key-value stores, and data formats like NetCDF and HDF5 are fundamental infrastructure components required by many professors in physics, biology, medicine, and computer science at the JGU and there is a huge demand for scalable big data and HPC platforms. The ZDV datacenter at the JGU, a tier 2 datacenter in Germany, will directly integrate successful techniques and tools developed in DIO into its production HPC environment and therefore offer its scientists direct access to this cutting-edge technology. The datacenter infrastructure will therefore provide a DIO reference platform, which helps datacenters in Germany and Europe to be able to evaluate DIO results in a production environment. ZDV will therefore, e.g., promote DIO results inside the German Gauss Alliance. ZDV will include Ph.D. and Master students into the development of the DIO architecture and Prof. Brinkmann will include selected topics from DIO into his lectures and seminars. The JGU and the ZDV furthermore always strives for academic excellence and will publish results achieved in this project at top systems conferences and journals like Usenix ATC and FAST, VLDB, IEEE/ACM Supercomputing, and/or ACM Transactions on Storage.
ICCS	ICCS is investing significant effort in the research work of WP5 and WP6. Through these activities, ICCS is targeting a significant improvement of its prior work on hardware accelerators for heterogeneous computing and storage systems. ICCS has identified this topic as one of the strong research outcomes of ICCS's Microlab group with significant potentials for exploitation, (either in the form of IPR management or through the creation of a start-up) and will support actively any future development with business plans and promotion activities. Finally, it is important for ICCS that DIO will support 1 post-doctoral and 2 Ph.D. researchers to work on FPGA-based acceleration technologies for the I/O path. Additionally, at least 3 Masters or undergraduate students are expected to get acquire significant insight in the area of the project by completing their undergraduate or Masters thesis in the context of DIO.
BULL	With respect to DIO, BULL has significant interest in architectural issues related to storage, analysis of overheads, and checkpointing. Analytical models extrapolated to exascale computing already predict a 30min system mean time between failures (SMTBF), thus implying the mandatory usage of non-volatile memory as intermediate step in a hierarchical checkpointing. In addition, checkpointing of hybrid architectures requires new methods and techniques to handle in a coherent manner

	heterogeneity and state on accelerators. Beyond, failure mitigation and fault tolerance, fast checkpointing and restart have many other potential uses in HPC for process migration, post-processing and debugging. BULL plans to design within DIO next generation checkpointing technology in conjunction with deep storage hierarchies and fast devices and to explore solutions for future architectures and products. In addition, as computing is looking to the next challenges of exascale computing for which data movement is part of the crucial optimization path, one of the main BULL's objectives is the exploitation of DIO results as "Proof of Concept" on existing HPC infrastructures augmented with next-generation of storage devices. Such PoC will facilitate a large spectrum of measurements on performance, scalability, and energy consumption to demonstrate benefits over existing solutions and will provide useful input to the roadmap of BULL towards exascale. Finally, BULL will consider the introduction of DIO technologies in its future products as a differentiating factor in the improvement of I/O performance and the ability to deliver reliable and Big Data enabled systems.
CYB	If Cybele is able to reach the expected performance thanks to the access to much larger datasets as enabled by DIO, CYB will provide a commercial service of agricultural production forecast with business models oriented toward farmers cooperatives and insurance companies. Although the application has been developed internally to the company as part of its R&D strategy, CYB has already established many contacts with industrial partners that are interested by the approach and the service. CYB will then have major interests in commercial solutions of exascale HPC infrastructures based on technologies provided by DIO that turn out to be crucial for its application.

Knowledge Management and Protection Strategy

For the success of DIO it is essential that all project partners agree on explicit rules concerning IP ownership, access rights to any Background and Foreground IP for the execution of the project and the protection of intellectual property rights (IPRs) and confidential information before the project starts. Therefore, aspects of innovation and management of knowledge and intellectual property are a significant part of the planned work in the respective work packages and will maximize the return from the research and protect the investments made.

In terms of managing the knowledge and intellectual property that will be generated, the consortium will use the following, well understood principles:

i. Generated knowledge is owned by the partner(s) that have produced it. This is a basic requirement as partners of the project have active interests in the technologies that will be generated. The partner generating a piece of knowledge will have a strong saying in how this will be used. There is a number of issues associated with joined intellectual property, however, partners have experience in dealing with these issues in both industrial as well as academic contexts and thus, will include more detailed guidelines and rules in a consortium agreement that will be signed before commencement of the project.

ii. Results of the project will be published in technical venues, outlining the new solutions and technologies developed. This is important for two reasons: (a) new technologies need to be subjected to comments of technical committees putting it to test in comparisons with other similar techniques; (b) new technologies need to generate traction and demonstrate their potential early on. For these reasons, we believe that basic results about the technology should be published at strong technical venues. Examples of such venues are given in the previous subsection.

   A concern that sometimes accompanies this principle is that some of the published information may be sensitive and of commercial interest. Although this is a valid concern, usually, information about base technologies that is published in technical venues is not at a level that allows immediate replication of the techniques by other groups. Moreover, the implementation of the proposed techniques is usually in these areas a challenge by itself, providing the team that has developed the technology with a substantial lead in the time-to-market. Naturally, if a new technology is published and is not pursued within a reasonable timeframe, it is possible for other teams to replicate the proposed techniques. However, in most cases of important technologies, these issues can be resolved upfront and before publication. The consortium agreement will layout guidelines to follow in the cases where the article contains confidential material, or where the technology is in the process of being patented.

iii. Sensitive pieces of information that may be crucial for the expected productization phase to follow this work will be kept as trade secrets by partners. It is possible that certain pieces of intellectual property can be patented, however, the usefulness of patents for critical technologies is usually limited, as legal procedures are time consuming and costly, especially when large companies are involved. Thus, partners will use their experience on which technologies should be patented and which ones will be kept as trade secrets.

It is important to note that each of these principles will rely on consensus and that the exact method to resolve possible conflicts will be described in the consortium agreement. As these principles have been used in the past successfully, we expect that, especially given the expertise of the consortium, they will form a sound foundation for managing the generated knowledge and intellectual property. As starting point for the consortium agreement we will use the MCARD-2020 agreement template model (provided by Digital Europe). Finally, dissemination and exploitation activities will also be aligned with the spirit and procedures of the consortium agreement.

Existing licensing restrictions

Next, it is important to clarify any issues with pre-existing IP with other components in the stack. DIO will develop a new storage stack (and tools) where components will have compatible licenses. In addition, the project will make available the DIO storage systems for research purposes in an open manner and with an appropriate license, as will be determined in the consortium agreement. To consider licensing restrictions with respect to external components, the major layers in the stack are at a high level:

Layers	License category
Applications	Various Licenses, Open and Proprietary
I/O Libraries and SLURM	Various Licenses, Open
Storage stack and tools	DIO partner proprietary as will be specified in the consortium agreement with standard interfaces to external components

At the bottom layer, the DIO storage stack in its implementation will interface with other components, e.g. memory allocators and the OS kernel with standard interfaces that do not impose significant restrictions[99]. In the middle layer, the I/O libraries that sits between DIO and applications includes components of varying licenses. However, these components are widely used and are open for use by the DIO project. In addition, DIO will share back any modifications to these libraries, as it is important to open interfaces to the community for broader use of its own technology.

Finally, partner applications in the project come with varying licenses and certain components are proprietary. In DIO the proprietary components are owned by partners, e.g. CYB, and therefore the project can make progress without licensing issues. The following table shows the components that will be offered to the project as pre-existing background and they can be used by partners for the purposes of the project.

Software Component	Owner	License category
UEABS	N/A	Open (Various licenses)
Cybele	CYB	Proprietary, open to project partners for the DIO project
NEMO	N/A	Open (CeCILL license)

As a summary, DIO does not have any significant unresolved license dependencies and restrictions and plans to produce new technology that will be available openly for research purposes by industry and academia.

2.4.3 Communication activities

Communication covers all actions by which activities, knowledge, results and ideas generated in the project are shared with the relevant stakeholders and via appropriate channels and mechanisms. This includes:

- High-level messages about technology limitations and solutions towards "executive-type" audiences.
- Technical messages and information at different levels for research/engineering audiences.
- Awareness messages towards the general public.
- Exploitation messages for business development audiences.

Although in all WPs contacts are made and networks are built, the focus of the communication actions will be grouped in WP7 in which all consortium members will participate. These activities will be on-going throughout the whole duration of the project and all partners are responsible for contributing with their achievements to this WP. For this purpose, WP7 will develop a dissemination and communication strategy. This strategy will be the basis for

[99] We foresee that DIO might have a small part in the Linux kernel for optimizing certain aspects of the non-common I/O path and this part will be under GPLv2 (or dual GPL/BSD) due to the Linux kernel restriction.

all communication activities and will ensure effective dissemination of the project's results and activities.

To be effective, our communication strategy will:

1. **Identify and refine target group audiences:** DIO will disseminate project results through a variety of measures, as indicated above and towards a variety of audiences. As part of our strategy we will identify the key audiences for each type of activity. To achieve this, we will start from partner collaborator networks and we will refine the categories based on on-going interests of individuals and organizations.
2. **Define the messages to communicate:** For each audience we will define a main message and associated material that needs to be communicated. These messages need to be at the right level, e.g. referring specifically to the technology, its features, or its benefits, depending on the audience.
3. **Select modes of communication for each audience:** There are various modes of communication and typically different audiences are used to expect different modes. The main modes of communication DIO will use are:
   - Leaflets that mostly present high-level messages.
   - Slide presentations that convey a complete story about the project at an appropriate level.
   - Demos that aim to show how the proposed technology addresses real needs in working systems.
   - Audio-visual material that aims to convey project goals in a concise and brief manner.
   - General-press (typically short) articles that generate awareness and raise interest.
   - Technical (typically detailed) articles that discuss specific problems and solutions in more detail.
   - Newsletters that provide an overview of activities and status.
4. **Identify the appropriate channel:** For each audience and mode of communication we will identify a set of channels to use, from the typical range of:
   - Project Website: We will create a project website with sections for different audiences.
   - Partner Websites: We will also use in a complementary manner the existing web pages of project partners to post appropriate messages and material.
   - Conferences, journals, workshops for more technical audiences and messages.
   - European and National projects and concertation events, for further details on DIO technology but also for facilitating synergies with other technologies as well, especially related to networking, processing, memory, and domain-specific applications.
   - Academic courses, theses, and conference/workshop tutorials for educational purposes.
   - General press for articles with broader appeal to less technical audiences. Project partners already have access to such channels and will start from existing contacts with press and journalists.
   - A mailing list (database) that we will create starting from partner collaborator networks and we will expand using an opt-in/opt-out model.
   - Social media for short updates and high level achievements for multiple audiences.

   DIO will pay particular attention to certain European-level tools for communication with the specific goal of developing further synergies at different levels:

   - Concertation events for European projects.
   - European technology platforms.
   - Roadmaps for technology and ICT research.
   - Public interest groups.
5. **Industry Day:** To facilitate several dissemination goals and towards the end of the project, DIO will organize an Industry Day at STFC, using an established practice of STFC to bring technology and applications together[97]. This event will focus on project technology and how it can address modern and future needs.

We expect that DIO activities will be successful in clearly and strongly communicating the main messages to different audiences, especially with respect to the fundamental technologies and their benefits for next-generation HPC and Cloud infrastructures.

3 Implementation

3.1 Work plan — Work packages, deliverables

DIO will use a simple WP structure that directly maps to the main technical and non-technical aspects of the project (Figure 13, Figure 14), as follows:

WP1 Project management (FORTH): WP1 will include all tasks related to project management and ensuring proper execution of all tasks in the project, as well their dependencies and interactions.

WP2 Data storage and access mechanisms (FORTH): WP2 will provide the core mechanisms related to data storage and data access based on an innovative key-value store that reduces overheads, fundamentally improves the way data and metadata are managed and accessed, addressing many of today's shortcomings over fast storage tiers, and dealing with reliability and availability.

WP3 Storage abstractions and I/O libraries (JGU): WP3 will examine issues related to providing legacy abstractions so that existing I/O libraries, including checkpointing, and applications can benefit from the proposed technology and will also examine new abstractions that can further improve application performance and system utilization, including data transformations, allowing larger datasets to be handled in a more interactive manner.

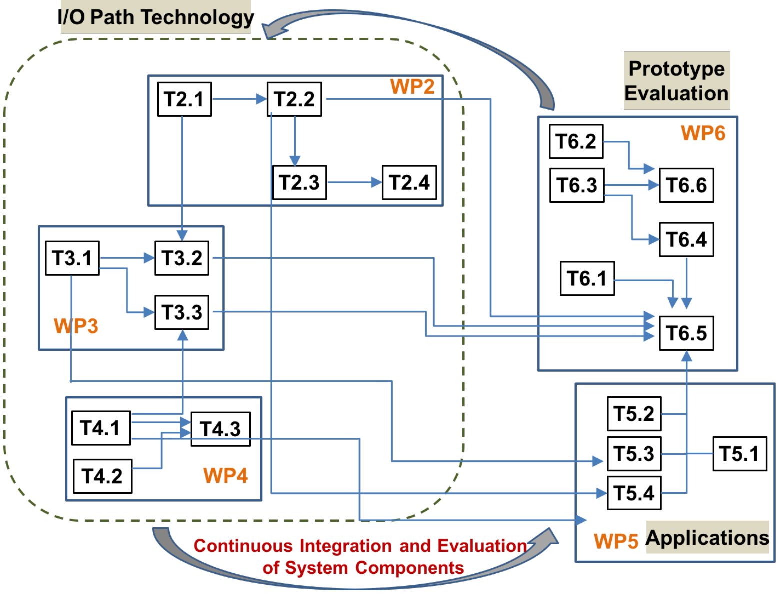


Figure 13: Pert chart showing how project WPs and components inter-relate.

WP4 Resource and QoS management (BSC): WP3 will examine issues related to ensuring high resource utilization while maintaining application QoS under high degrees of concurrency and interference across applications, over heterogeneous storage tiers, including fast devices. WP4 will provide the monitoring and policy aspects to ensure that future systems operate efficiently and applications see the expected benefits under dynamic conditions.

WP5 Application adaptation and evolution (STFC): WP5 will work on understanding modern workloads in terms of the I/O behavior and on integrating the project applications and use-cases (Environmental modeling, Procurement benchmarks, Checkpointing) with the prototype software stack. Equally important, WP5 will also examine how applications should evolve in the future to best take advantage of emerging storage devices and the new systems software stack for performing I/O.

WP6 Acceleration, integration, and evaluation of the converged scalable architecture (ICCS): WP6 will examine how storage and FPGA-based acceleration should be incorporated in future racks so we can increase the efficiency and density of I/O. In addition, WP6 will evaluate the system prototype and will examine how future systems should be architected to best benefit from new device technologies and the proposed software stack. Finally, WP6 will examine issues related to scalability to multiple racks and how I/O performance can scale with system size.

WP7 Dissemination and exploitation (BULL): WP7 will orchestrate and execute communication activities and actions related to dissemination and exploitation of the technology DIO develop, as well as interactions with the end-to-end chain to raise awareness and ensure impact.

Next, in Table 3.1a we discuss in more detail each WP, in Table 3.1b we summarize WP effort, and in Table 3.1c we summarize all deliverables.

MILESTONES												MS1												MS2												MS3
Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
WP1: Project Management [lead: FORTH]																																				
T1.1: Project organization and management	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T1.2: Project quality management	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Deliverables	**D1.1**		**D1.2**									**D1.3 (PPR)**												**D1.4 (PPR)**												**D1.5 (PPR), D1.6**
WP2: Data storage and access mechanisms [lead: FORTH]																																				
T2.1: Single-node, multi-tier, memory-mapped, KV-based data access	■	■	■	■	■	■	■	■	■	■	■	■																								
T2.2: Multi-node, scale-out, KV-based data access													■	■	■	■	■	■	■	■	■	■	■	■												
T2.3: RDMA-based client access and data replication over fast devices																			■	■	■	■	■	■	■	■	■	■	■	■						
T2.4: Optimization for transparent and explicit data reorganization and migration across devices, tiers, and nodes																									■	■	■	■	■	■	■	■	■	■	■	■
Deliverables												**D2.1**												**D2.2, D2.3**												**D2.4**
WP3: Storage abstractions and transformations [lead: JGU]																																				
T3.1: I/O library support	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■												
T3.2: Checkpointing library over new system for existing and emerging applications													■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T3.3: New storage abstractions and APIs for emerging data processing applications in HPC													■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Deliverables						**D3.1, D3.2**						**D3.3, D3.4**												**D3.5**												**D3.6, D3.7**
WP4: Resource and QoS management [lead: BSC]																																				
T4.1: Design and implementation of I/O and accelerator profiling	■	■	■	■	■	■	■	■	■	■	■	■																								
T4.2: Dynamic resource allocation for mixed and highly concurrent workloads	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■												
T4.3: Policies for transparent data reorganization (placement and migration)													■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Deliverables												**D4.1, D4.2**												**D4.3**												**D4.4**
WP5: Application adaptation and evolution [lead: STFC]																																				
T5.1: UEABS benchmarks analysis of I/O usage and skeleton benchmarks for studying I/O characteristics in depth													■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■						
T5.2: NEMO/XIOS workflow adaptation	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T5.3: Code_Saturne workflow in a fuel assembly large-scale application adaptation	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T5.4: Crop identification and growth modeling workflow adaptation	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Deliverables												**D5.1[a,b,c]**												**D5.2[a,b,c], D5.3**												**D5.4[a,b,c]**
WP6: Acceleration, integration, and evaluation of the converged scalable architecture [lead: ICCS]																																				
T6.1: Prototype for continuous integration and evaluation of system components	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T6.2: High-level fully functional architectural simulator	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■																		
T6.3: Design space exploration of data transformations with respect to accelerator, storage, and processing alternatives	■	■	■	■	■	■	■	■	■	■	■	■																								
T6.4: FPGA accelerator for cut-through data processing and layout transformations								■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■												
T6.5: Evaluation of integrated prototype including I/O and acceleration and the software stack																									■	■	■	■	■	■	■	■	■	■	■	■
T6.6: Extrapolation to large scale systems																									■	■	■	■	■	■	■	■	■	■	■	■
Deliverables												**D6.1, D6.3**						**D6.2**						**D6.4**												**D6.5**
WP7: Dissemination and Exploitation [lead: BULL]																																				
T7.1: Dissemination and communication strategy and plan, and Data Management Plan	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T7.2: Dissemination and communication material	■	■	■	■	■	■																														
T7.3: Dissemination and communication actions	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
T7.4: Management of IPR, Exploitation and Commercialization							■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Deliverables			**D7.1, D7.2a**			**D7.3**						**D7.2b, D7.4a, D7.5a**												**D7.2c, D7.4b, D7.5b**												**D7.2d, D7.4c, D7.5c,D7.6**

Figure 14: Gantt chart with the timing of work packages and their tasks, deliverables, and milestones.

Table 3.1a: Work package description

WP1	Lead beneficiary			**FORTH**			
Work package title	Project Management						
Participant number	1	2	3	4	5	6	7
Short name of participant	**FORTH**	BSC	STFC	BULL	JGU	ICCS	CYB
Person/months per participant	12	2	2	4	2	2	2
Start month	1		End month		36		

Objectives

This WP will be devoted to the following objectives:

- Setup and maintain the project internal communication tools for management and monitoring purposes.
- Manage and administer technical and financial aspects of the project at the WP, activity, and partner levels, using milestones and risk analysis.
- Compile the annual management and progress reports, including financial aspects of the project.
- Interface with the EC for reporting purposes and any issues that arise.
- Ensure the timeliness and high quality execution of deliverables and other activities of the project.

Description of work

Task 1.2 Project organization and management [M1-M36] [FORTH]

This WP will conduct all project management related activities. Management activities will be centered around Management Board meetings, project milestones, and risk analysis. The Management Board will meet periodically (at least quarterly), both physically during technical meetings and via conference facilities.

The Technical Coordinator will use conventional Project Management tools to help run the entire operation. A Quarterly Technical Meeting of all partners will focus on achievement of deliverables and updates of the normal management information (adherence to time scales, budgets and specifications). In order to provide regular but independent assessment and evaluation of progress, the consortium will use an internal review procedure for reports, to both determine the progress achieved and suggest corrective actions. In addition, this task will continuously monitor progress based on Milestones and Risks and will revise each category, including updates to periodic reports.

The Project Manager will maintain a schedule of partner obligations and will send out regular reminders of reports, deliverables, Cost Statements, etc. coming due, and will co-ordinate with the EC Project Officer as required. The Project Manager will also co-ordinate with the WP7 leader the dissemination activities; moreover, the Project Manager will co-ordinate arrangements between partners to extend their co-operation into the industrialisation and exploitation phase, starting with the Consortium Agreement and IPR management issues.

Task 1.2 Project quality management [M1-M36] [FORTH]

This task is responsible for ensuring quality across documents, actions, and all communication performed by the project internally and externally. The task will produce early in the project timeline a quality plan to specify the quality procedures, quality reviews, peer reviews, control, monitoring and reporting activities which will ensure that the prerequisite quality standards are met. The task will ensure that information will flow internally across teams, WPs, tasks at a high quality so that teams and team members can be productive and are not delayed by the lack of proper communication. More importantly, this task will apply quality control to all external communication including deliverables, social media, press publications, technical publications, audio-visual material, and web-page content. The task will define the procedures for quality control at the beginning of the project, in a manner that does not hinder fast progress and adapting to changing conditions and which suites the work style of all partners as well as the type of work being conducted in the project, i.e. rapid prototyping in multiple cycles and continuous integration of technologies to a main prototype.

Deliverables

D1.1 Public project presentation **[M01] [FORTH]**
D1.2 Quality management plan **[M03] [FORTH]**
D1.3 First periodic report **[M12] [FORTH]**
D1.4 Second periodic report **[M24] [FORTH]**
D1.5 Third periodic report **[M36] [FORTH]**
D1.6 Final project report **[M36] [FORTH]**

WP2	Lead beneficiary			**FORTH**		
Work package title	Data storage and access mechanisms					
Participant number	1	2	3	4	5	6	7
Short name of participant	**FORTH**	BSC	STFC	BULL	JGU	ICCS	CYB
Person/months per participant	54	13	13		12		
Start month	1		End month		36		

Objectives

The main objective of this WP is to design and implement the key-value store mechanisms required for storing and accessing data and metadata in an efficient manner and which allows the system to deliver the performance of modern storage devices while achieving high resource utilization.

- Development of a single node, low-overhead KV store
- Development of a scale-out, low-overhead KV store
- Fault tolerance and availability functionality in the KV store
- Load balancing and dynamic adaptation of the KV store

Description of work

Task 2.1 Single-node, multi-tier, memory-mapped, KV-based data access [M1-M12] [FORTH, JGU, BSC]

This task will create a highly efficient write-optimized key-value store for fast storage devices (SSDs, NVM, etc.) that stores data in a versatile form of (key, value) pairs and which will interact with data (and appropriate metadata) through a simple directory API consisting of functions *get, put, scan, delete keys*, together with a small number of management functions to *create, remove, split, and merge key regions*. Unlike previous work, this task will focus on fast storage devices that provide a high potential for reducing CPU overhead by allowing some level of random accesses without any significant impact on throughput (unlike disks). In this context, this task will:

- Organize metadata in an index that operates efficiently over fast devices and allows the system to maximize device utilization. We will leverage a multi tiered storage system that will consist of DRAM, NVM, SSD, emerging SCM, and a parallel filesystem over hard drives (as the last slow tier). To reduce overheads, *DIO* will not operate on sorted buffers (as most current systems do), but instead, it will maintain a B-tree index within each level of the KV store. As a result, this approach generates smaller I/O requests in favor of reduced I/O and reduced CPU overhead.
- Use memory mapped I/O, unlike all current approaches that use explicit I/O, which has the potential to remove OS kernel overheads and expensive I/O cache lookups. Memory mapping uses a single address space for both memory and storage and virtual memory protection to determine the location (memory or storage) of an item. This eliminates the need for pointer translation, which occurs during index operations regardless of whether items are in memory or not.
- Manage fast storage devices directly for allocation, recovery, and synchronization purposes, without any intervening layers that introduce overheads.
- Finally, as discussed, the KV store will optimize CPU cycles per operation, which is the most important metric going forward, given current limitations on power and the required system sizes that we need to achieve.

Task 2.2 Multi-node, scale-out, KV-based data access [M13-M24] [FORTH, JGU, BSC]

This task will extend the KV store to operate over multiple nodes. The challenge of the task is to assign different keys to compute nodes. The main issues to deal with in this task are:

- Introduce metadata and the associated mechanisms for identifying the location of data for *put, get, scan* operations. Essentially, this requires two levels of indexes, one intra-node and one inter-node. We will examine approaches where the inter-node metadata for location purposes can be maintained in memory for efficiency and can be replicated widely to avoid network communication.
- Introduce metadata and the associated mechanisms for adding and removing nodes dynamically as the system grows. This will require keeping track of node membership and the ability to perform asynchronous notifications when system membership changes, in a consistent, system-wide manner.
- Mechanisms for splitting and merging key regions across nodes. This is important for dynamically growing and shrinking the system as data size changes over time.

All performance-critical network operations will use RDMA-based communication to reduce CPU overhead in the common path.

Task 2.3 RDMA-based client access and data replication over fast devices [M19-M30] [STFC, FORTH]

This task will extend the scale-out KV store of T2.2 to use data replication to address two fundamental challenges of distributed key-value stores: fault tolerance and service availability. This task will address the main issues that such a replication creates: 1) the slowdown of writes, and 2) space explosion.

First, given that data replication in a consistent way slows down writes, the task will implement data replication using RDMA directly between user-space of the replicas and the buffers of the nodes that hold the primary copies of the data. Thus, replication will bypass completely the slower devices and network protocols and will instead occur directly between the faster storage tiers of the hierarchy, using RDMA.

Second, the task will reduce the space overhead of data replication through customized coding techniques that result in lower overhead and simpler data management protocols. Coding techniques, such as erasure coding, have the potential to reduce the space overhead of fast devices. However, current approaches to coding tend to incur high performance overheads. This task will explore the potential of recent developments in coding techniques and protocols to achieve better space utilization without introducing prohibitive performance overheads.

Task 2.4 Optimizations for transparent and explicit data reorganization and migration across devices, tiers, and nodes [M25-36] [FORTH, STFC, BSC, JGU]

This task will provide optimizations to the scale-out KV store of T2.2. More specifically, we will examine:

- Efficient mechanisms for load balancing of data across devices, tiers, and nodes. Given that modern workloads are dynamic and system infrastructure is becoming complex because of size and diversity, it is important to incorporate in the system efficient mechanisms for migration of key-ranges based on use and policies, in collaboration with WP4. These mechanisms (and associated polices) will allow the system to adapt to workload changes. In particular, it is important to examine efficient data movement between nodes connected by appropriate networking using remote direct memory access (RDMA). It will provide an API that will be used by the DIO scheduler layer to transparently invoke copying of data directly to node that requires it, instead of passing through the operating system stack.
- Efficient mechanisms for splitting and merging key-ranges without unnecessary movement of data on devices (and across nodes). Typically, splitting and merging requires reorganization of data on devices, which is expensive and results in significant drop in the performance perceived by applications because the devices are monopolized by data movement. However, in multi nodes setups there are opportunities to move data across nodes without generating excessive local device traffic.
- Efficient mechanisms for identifying and optimizing access to hot data so that certain nodes do not become a bottleneck.

Overall, this task will exploit the indexing mechanisms of the KV store to minimize data movement between storage tiers and across nodes during split and merge operations, and to provide transfers across nodes at the same time as data is being (re)placed on devices, thus avoiding significant overheads of today's systems and approaches.

Deliverables

D2.1 Design and implementation of the memory mapped key-value store **[M12] [FORTH]**
D2.2 Design and implementation of the scale-out key-value store **[M24] [FORTH]**
D2.3 RDMA-based replication and data transfer mechanisms for data reorganization **[M24] [STFC]**
D2.4 Final prototype of optimized key-value store **[M36] [FORTH]**

WP3	Lead beneficiary			**JGU**		
Work package title	Storage abstractions and I/O libraries					
Participant number	1	2	3	4	5	6	7
Short name of participant	FORTH	BSC	STFC	BULL	**JGU**	ICCS	CYB
Person/months per participant	21	12	4	21	30	6	7
Start month	1		End month		36		

Task Objective

- Integration of legacy I/O libraries into the DIO stack in order to provide legacy applications with access to DIO
- Design and implementation of appropriate support for emerging HPC applications
- Design and implementation of multi-level checkpointing libraries over the DIO storage API

Description of work

We will port HDF5 and MPI-IO in task T3.1, which will also provide indirect support for NetCDF that interoperates with HDF5. Popular checkpointing libraries will be integrated in task T3.2. Direct support for emerging HPC applications will be investigated in task T3.3.

Task 3.1 I/O library support [M1–M24][JGU, STFC, CYB, BSC, BULL]

I/O middleware libraries have the advantage that datasets are (mostly) stored based on a self-describing approach, which simplifies the exchange of data between applications and datacenters and removes, e.g., endianness problems. HDF5 structures datasets and can present them as different views, so that data transformations are performed by HDF5 without requiring application involvement, removing unnecessary burden from the programmer and removing subtle errors, which are difficult to debug. T3.1 will directly integrate DIO support into HDF5, which (typically) stores data items belonging together in files. Nevertheless, its Virtual Object Layer (VOL) allows plugins to interface storage APIs different from POSIX, like DAOS, where HDF5's datasets themselves can be stored as simple arrays, fitting to KV pairs with large values, while attributes and groups can be stored as KV pairs with small values[100].

Another important file-based middleware is MPI-IO[101], which allows file system aggregator nodes to derive an optimal data layout of many nodes writing to the same file applying a multi-phase negotiation protocol. Thus, in data writes, first compute nodes transfer data to aggregator nodes, which in turn persist them to the storage backend. The MPI-IO consistency semantics is very precise, but less strict than POSIX[102], which makes it a good candidate to be implemented on top of DIO. T3.1 will therefore architect and implement an MPI-IO interface layer to DIO, which leverages the fast-path to the underlying storage systems, which will support the most important MPI-IO consistency semantics, especially used in checkpointing applications.

Task 3.2 Checkpointing library over new system for existing and emerging applications [M13–M36][BULL, FORTH, JGU]

The task will design and implement direct support for a multi-level checkpointing library to support legacy applications that are not willing to modify their code. The checkpointing library will leverage DIO's ability to inherently cope with multiple tiers of storage transparently or via its explicit API. One of the goals in T3.2 is to explore which approach is more appropriate in the case of state checkpointing.

Legacy applications will be covered by providing multi-level checkpoint capability for MPI-IO, HDF5, and transparent checkpointing systems, such as DMTCP. A challenge is that the information of which stored data constitute parts of a checkpoint is often known only to applications. In the absence of such information the DIO library cannot distinguish checkpoint data and thus treat it differently than ordinary storage entities in the I/O path. Therefore, DIO will treat checkpoint data of legacy applications similar to other stored entities, deciding implicitly where it is placed in the storage hierarchy, aiming for a configurable level of failure resiliency. Recovery to a recent consistent state will be managed by the application performing the checkpointing.

By supporting emerging APIs for multi-level checkpointing, such as SCR and FTI, DIO will support a wide range of modern applications that have been ported or will be ported to use these interfaces. Since these interfaces are straightforward and developed to allow an easy migration path for applications, this approach is in line with prevailing trends in the HPC community.

When applications identify checkpoint data and operations to DIO through explicit calls (e.g., that an upcoming creation of a data entity and/or write to it is a checkpoint, its desired level of fault tolerance, etc.), DIO will be able to operate in an improved mode by organizing data appropriately and by mapping checkpoint operations to DIO data-snapshotting copy-on-write (CoW) operations which can freeze an entire (checkpoint) dataset as of a specific point-in-time. The dictionary API of DIO allows the checkpointing library to use simple mechanisms, such as tags and key-ranges to specify checkpoint sections and how they should be placed in specific storage tiers (including local and remote memory). The key-value approach also enables applications to perform deduplication of checkpoints within the data path[103]. The DIO checkpointing library will use the APIs and data movement mechanisms to be provided in WP2 to the management of checkpoint data of HPC (and cloud) applications throughout their lifecycle (from creation to obsolescence via creation of a subsequent checkpoint). DIO will be able

[100] Lofstead, J.F., Jimenez, I., Maltzahn, C., Koziol, Q., Bent, J., Barton, E.: *DAOS and friends: a proposal for an exascale storage system*. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2016, pp. 585-596, 2016

[101] Gropp, W. D., Lusk, E. L., Skjellum, A.: *Using MPI - portable parallel programming with the message-parsing interface*. MIT Press 1994, ISBN 978-0-262-57104-3

[102] Gropp, W. D., Lusk, E. L., Sterling, T. (editors): Beowulf Cluster Computing with Linux, Second Edition, The MIT Press, ISBN 0-262-69292-9

[103] Kaiser, J., Gad, R., Süß, T., Padua, F., Nagel, L., Brinkmann, A.: *Deduplication Potential of HPC Applications' Checkpoints*. In Proceedings of the 2016 IEEE International Conference on Cluster Computing (CLUSTER), pp. 413-422, 2016

to remove obsolete checkpoints as soon as more recent checkpoints have been successfully created, taking inter-dependencies between different checkpoints into account. By virtue of DIO's global I/O address space, recovery will be feasible from a different set of compute-nodes than in the original (pre-failure) execution.

This task will furthermore carry out the design and evaluation of new checkpointing protocols to handle heterogeneous environments using computing accelerators as co-processors, whose state must also be checkpointed.

Task 3.3 New storage APIs for emerging HPC applications [M13–M36] [JGU, BSC, STFC, CYB, FORTH, ICCS, BULL**]**

First, this task will examine how high-level representations of data can be used to automatically perform data transformations in the I/O path. Emerging applications, such as the climate and environmental modeling library XIOS, include a precise, self-explaining XML-description of the stored data types as an additional recipe besides the main data. They typically use the XML-description (including hints from the application developer) to re-arrange data accesses in a write-friendly manner, e.g., as logs and can therefore substantially improve (asynchronous) write performance. Furthermore, they can include filters describing data transformations, so that not all data has to be stored, but some of it can be compressed (leaving out unimportant data) or regenerated on the fly. In T2.2 the DIO key-value interface will be coupled with the XIOS library to show the benefits of the proposed data access mechanisms.

Second, this task will examine opportunities for customizing data interfaces and mechanisms for different granularities. Data within I/O libraries and applications can be distinguished into smaller key-values like measuring units and bigger objects, e.g., data matrices. T3.3 will couple the advantages of key-value stores and object stores for emerging HPC applications by building additionally to the DIO key-value store an object storage API that can scale to more than existing systems, as each object can be managed independently and within the application address space. DIO provides a natural API to build such object stores, as the object names can be interpreted as keys and the (byte-stream) object data as values. T3.3 will therefore use the DIO-API to build an object-like storage system, which is closely compatible to the Amazon / Swift-API and which supports an extended metadata handling for each object. In contrast to existing object stores, it can also use the fast LSM object lookup path to efficiently manage and access many small objects.

Deliverables (brief description and month of delivery)
D3.1 Requirements of applications and standard I/O libraries over DIO **[M6] [JGU]**
D3.2 Design of DIO multi-level checkpointing library for heterogeneous architectures **[M6] [BULL]**
D3.3 First prototype of DIO multi-level checkpointing library **[M12] [FORTH]**
D3.4 First adaptation of standard I/O libraries over DIO **[M12] [JGU]**
D3.5 Full implementation of DIO multi-level checkpointing library **[M24] [FORTH]**
D3.6 Novel APIs for data transfer and transformation in systems with deep storage hierarchies **[M24] [JGU]**
D3.7 Final adaptation of standard I/O libraries and novel APIs over DIO **[M36] [JGU]**

WP4	Lead beneficiary			**BSC**			
Work package title	Resource and QoS management						
Participant number	1	2	3	4	5	6	7
Short name of participant	FORTH	**BSC**	STFC	BULL	JGU	ICCS	CYB
Person/months per participant	12	48	12		6	6	
Start month	1		End month		36		

Objectives

The objectives of WP4 are:

- Extend the SLURM scheduler (used extensively today) to gather I/O QoS requirements from applications.
- Create a monitoring/profiling system for storage and network I/O, at the I/O operation granularity.
- Use this system with QoS requirements to estimate the required I/O resources for each application.
- Enhance SLURM with a *new storage scheduler* that allocates resources to concurrent applications and processes and also colocates applications based on QoS.
- Use the new storage scheduler to *dynamically* increase or decrease the amount of data resources used to satisfy application QoS and to ensure that I/O resources are efficiently utilized.

Description of work

Task 4.1: Design and implementation of I/O and accelerator profiling [M1-M12][STFC, BSC, ICCS]

In order to understand how to improve application I/O, it is necessary to first detect all I/O bottlenecks. This task will develop a user-friendly tool that works across different applications to provide insights behind I/O bottlenecks. The tool will operate in two modes. 1) A non-intrusive mode, where the tool collects I/O information dynamically without user intervention. 2) An intrusive mode that will provide an interface to users to statically instrument applications with hints that help with collection of statistics during application execution. Both modes will provide information at the Region of Interest (ROI) level, which is the finest possible granularity, such as the line of code or function name. In the absence of related monitoring technology today, we will also incorporate statistics for the use of accelerators that WP6 will add to the storage path.

To achieve this goal, our tool will work on top of commands such as *blktrace/blkparse,* micro-architectural hardware events such as cache misses, and tools that provide statistics for the accelerators. For efficiency reasons, we will combine *statistical even sampling* with applications' stack trace to associate ROIs with I/O blocks. The tool will advance the current state of the art with the addition of a wide range of granularity levels. All I/O blocks will be tagged and monitored from the moment they leave the lowest storage tier. Using this information we will be able to produce metrics such as: 1) the amount of time an I/O block has been accessed for computation (*temporal locality)*, 2) how often neighbouring I/O blocks are accessed (*spatial locality),* 3) whether data access is contiguous or non-contiguous on the device (amenable for parallel I/O access), 4) determining the size of "hot" data (*I/O histogram*), 5) visual representation that shows where "hot" data resides (*I/O heat map)*, 6) visualizing access patterns over time (*I/O pattern map)*, 7) estimation of energy for I/O workflows, and utilization of the accelerators, data transfers to/from accelerators.

With this wealth of data generated by the monitor we will be able to identify where I/O bottlenecks are in the application and deduce what would be the best DIO techniques to use for removing these bottlenecks.

Task 4.2 Dynamic resource allocation for mixed and highly concurrent workloads [M1-M24] [BSC, FORTH, JGU]

The main purpose and challenge today for I/O resources allocation is to achieve application-level QoS while ensuring high resource utilization. For this purpose, we will extend a production scheduler, SLURM, with a dynamic resource allocation module that will combine QoS requirements of applications with system status metrics from T4.1. Initially, the application will provide I/O related QoS requirements, either in form of latency, or bandwidth, and the module will allocate the minimal set of resources that can meet them. Of course, as the system is used by different applications, the number of offered resources may change dynamically in order to reduce interference across large numbers of concurrent entities (applications, processes, threads, etc).

One of the main challenges is that often application performance does not improve as offered resources increase. For this reason, the module will adapt dynamically, as it will try different candidate allocations before choosing the most rewarding one. Our approach will use learning techniques that are appropriate for complex setups with diverse device technologies and large numbers of nodes, devices, and processes.

SLURM currently makes only static decisions when scheduling jobs that are grossly approximate and result either in low application QoS (when underprovisioning) or low resource utilization (when overprovisioning). Our scheduler, implemented as a plugin, will extent SLURM with dynamic decisions and the ability to adapt to varying workloads, a main challenge in today's infrastructures. The dynamic decisions of our scheduler will be conveyed to SLURM for modifying job reservations; our new storage scheduler will inform the application to adapt to the new environment with modified I/O resources.

For example, consider the resource allocation for two jobs; Job A requires low latency I/O operations at high concurrency, while Job B requires high throughput, but for a short period of time, e.g. when checkpointing. The tool of T4.1 will collect performance metrics for each job and convey the information to the modified SLURM, which will decide the minimal number of resources that can achieve the desired performance. Job A will receive a large amount of resources to maintain the low latency (because it is expected to perform a lot of operations) and Job B will receive a smaller set.

During the execution of both jobs, their QoS will be monitored. If they are acceptable, the modified SLURM will deallocate I/O resources to create room for more applications. If, on the other hand, the QoS is violated, our scheduler will increase the amount of I/O resources for the affected job, following a similar approach. Due to the infrastructure complexity and the workload diversity, our learning approach will be based on trial-and-error. Of course, the implementation will consider implications of each trial, such as overheads of the learning algorithm.

Finally, the storage scheduler will be able to track operations and prioritize operations from a job. Given that centralized approaches do not scale, these actions need to be taken in a distributed manner and at an appropriate unit. Therefore, the modified SLURM will operate on chunks of operation, rather than scheduling individual operations.

Overall this task will:

- Extend SLURM to use and communicate QoS Requirements.
- Create a new I/O scheduler plugin for SLURM to translate QoS and monitoring information into resource requirements using learning techniques and dynamic resource allocation actions.
- Create the infrastructure required for translating dynamic storage scheduler actions to application changes for adapting the use of I/O resources.

Task 4.3 Policies for transparent data reorganization (placement and migration) [M13-M36] [BSC, FORTH, JGU]

This task will examine the challenging issue of data reorganization for QoS and resource utilization purposes. The storage scheduler in SLURM will also be able to monitor data location and take decisions for data migration when necessary, either due to resource use or due to application performance. Although data placement has several aspects, in this task we will focus on data placement in different tiers of the deep storage hierarchy, which is a main challenge for increasing I/O performance in future systems.

We will implement different algorithms that allow the movement and redistribution of data across tiers, while considering the available I/O resources. Starting with a simple redistribution, we will investigate policies based on access patterns, performance of the devices, and available parallelism (from T4.1) to meet job QoS (from T4.2).

Our APIs will also include support for frameworks that perform their own scheduling, e.g. based on locality. First we will enable queries to the storage system about data location. Second, the APIs will allow frameworks to explicitly place data in the different layers of DIO hierarchy.

Overall, this task will:

- Interface with the storage scheduler agent to allow automatic decisions for data movement.
- Create mechanisms and policies that migrate data for optimal application QoS and resource utilization.
- Provide APIs for explicitly identifying and specifying data location.

Deliverables

D4.1 Design and implementation of non-intrusive tool for I/O and accelerator profiling. **[M12] [STFC]**
D4.2 Design and preliminary implementation of storage scheduler. **[M12] [BSC]**
D4.3 Complete implementation of the storage scheduler for dynamic resource allocation under high concurrency. **[M24] [BSC]**
D4.4 Design and implementation of scheduler-based data reorganization techniques. **[M36] [BSC]**

WP5	Lead beneficiary			**STFC**			
Work package title	Application adaptation and evolution						
Participant number	1	2	3	4	5	6	7
Short name of participant	FORTH	BSC	**STFC**	BULL	JGU	ICCS	CYB
Person/months per participant	3	24	24		9	6	18
Start month	1		End month		36		

Objectives

- In-depth analysis of I/O profile of UEABS benchmarks and generation of micro-benchmarks that reflect the obtained I/O profiles.
- Development of best-practices on how to improve I/O on benchmarks as a result of WP2 to WP5
- Adaptation of applications for the DIO stack: XIOS/NEMO, Code_Saturne, and Cybele
- Projections for new application capabilities, based on the improved DIO software stack

Description of work

This WP will develop the necessary benchmarks and applications to evaluate in-depth all technologies in the project. We will use the I/O profiler from T4.1 to analyze in T5.1 in-depth I/O characteristics of the UEABS benchmark suite and then create a set of skeleton benchmarks in T5.2 that will reflect the I/O behavior of UEABS. Tasks T5.3, T5.4, and T5.5 will adapt real-world applications NEMO/XIOS, Code_Saturne, and Cybele by referencing to best practices from task T5.2. These benchmarks and applications will then be used to develop the best-practices on how to adapt I/O part of the applications and to evaluate project technology.

Task 5.1 UEABS benchmarks analysis of I/O usage and skeleton benchmarks for studying I/O

characteristics in depth [M13-M30][STFC, JGU, BSC]

This task will use the profiling tools from T4.1 to analyze the Unified European Applications Benchmark Suite (UEABS), which STFC currently uses for procurement purposes. This suite consists of 12 programs: ALYA, Code_Saturne, CP2K, GADGET, GENE, GPAW, GROMACS, NAMD, NEMO, QCD, Quantum Espresso, SPECFEM3D. The objective behind this benchmark suite is to provide highly realistic and currently relevant codes and data sets that are executed on large Tier-0 and Tier-1 systems. By profiling this set of benchmarks, we will obtain insights on how these applications utilize the I/O path. Not only will these insights provide valuable feedback to the technical work packages regarding the I/O behavior of real workloads, but *as far as we are aware, this will be the first full, in-depth study of all 12 codes from UEABS benchmark suite at one place on utilization of contemporary I/O interface and identification of I/O bottlenecks in those codes.*

A limitation of the UEABS benchmark is that it cannot be used for micro-benchmarking during the development of DIO, because it only runs and measures its applications end-to-end. But in order to understand the impact of design decisions that were made as part of DIO work we want to be able to study in-depth I/O characteristics. Therefore, this task will also use the in-depth I/O analysis of UEABS benchmarks to create a set of representative *skeleton benchmarks* that will be used for micro-benchmarking of various DIO components. After analyzing DIO with this set of benchmarks we will be in a position to exactly pinpoint where the main differences are in performance and energy efficiency between contemporary I/O and DIO architecture. These benchmarks will be an invaluable tool not only for design exploration, but also for future work in HPC storage and I/O.

Task 5.2 NEMO/XIOS workflow adaptation [M1-M24][BSC, JGU, FORTH, ICCS**]**

This task will adapt an Earth-climate model application, NEMO, to DIO to fully take advantage of the potential of the proposed I/O stack. Currently, the I/O is managed by the XIOS I/O server (coded in C++), which uses NetCDF files to provide the output and the diagnostics generation of the ocean model. Other applications using XIOS will also benefit from work in this task.

The main goal of this task is to develop an interface and any extensions required, for XIOS to work in an optimized manner with DIO and perform production runs for technology evaluation purposes. This interface will deal with configuration files, and will manage all initial files. Furthermore, a new checkpoint method will be developed using DIO's approach. Finally, the task will examine what XIOS operations and NEMO functions can map to accelerators of WP6.

All these developments will be tested and benchmarked using current production code versions and configurations and extending configurations to produce high volumes of output, as the community projects problem sizes with grow. For instance, one of the cases that is important to explore is to deal with hourly 3D variables, enabling user analysis of large datasets practically in real time.

Task 5.3 Code_Saturne workflow in a fuel assembly large-scale application adaption [M1-M24] [STFC, JGU, FORTH**]**

Currently, to work on huge data set Code_Saturne breaks it down in smaller chunks on which it can efficiently operate. Whilst this works fine for mesh, domain partition and post-processing it is not suitable for restart file and checkpoint file when as/if the whole flow field is needed.

In order to enable Code_Saturne to work fully on huge data sets, this task will replace the application's MPI-IO interface that is used for distributed block-based access to fully take advantage of the capabilities of the new I/O stack. For this workload, DIO will follow a dual approach: On one hand we will examine the performance of Code_Saturne using the MPI-IO implementation of task T3.1, and the profiling tool from T4.1. We will also examine, in parallel with porting Code_Saturne's block-based I/O to DIO, how Code_Saturne might evolve to better take advantages of additional DIO features, such as new checkpointing library. We will decouple I/O from different processes by using private ranges (over a global namespace), that will then be implemented on top of DIO both natively and using intermediate I/O libraries.

The task will provide to Code_Saturne a range-based I/O library that will benefit from DIO's indexing mechanism and support for large numbers of ranges and concurrent writes. For range-based I/O we will use global numbering and we will explore important parameters, such as the role of the range size on performance and scalability This increased performance and scalability will enable Code_Saturne to work much more efficiently with restart and checkpoint files. On top of this, the use of the new checkpoint method developed by DIO will enable coupling of simulation execution with visualization process in order to visualize the results during the simulation. Finally, after adapting Code_Saturne to work with DIO we will be able to run large mesh sizes, which will allow researches to obtain better insight of the flow in fuel assemblies.

Task 5.4 Crop identification and growth modeling workflow adaptation [M1-M24] [CYB, JGU, FORTH**]**

In this task, we will adapt the workflow of the Cybele agricultural production forecast application to use DIO:

(a) Phase 1 (crop identification): Improved learning of the machine learning techniques with training over much larger areas thanks to the improved I/O, which currently is the main bottleneck in training. This is a use case for testing optimized I/O over geospatial data. In the application, I/O operations are managed by a Raster file driver written in C bundled with the open-source geospatial library GDAL. In a second step, the model will run identification for large-region queries (over the learned model) at timescales that allow near-interactive use of the system.

(b) Phase 2 (data assimilation for plant growth modeling and yield simulation): Improved accuracy for the models to achieve higher precision predictions for the evolution of plants thanks to data assimilation. Improvements in the workflow imply optimization of read/write operations of particle states to storage so that a much larger number of particles can be simulated. The I/O operations will be managed by the application through binary formats provided by ROOT or AVRO.

In addition, this task will examine how large numbers of application instances can run concurrently. This is extremely interesting for designing services that can be used to help different actors (scientists, policy makers, etc) to extract information and predict future actions in emerging domains, such as precision agriculture.

Deliverables

D5.1(a,b,c) Design of required adaption for each application (XIOS/NEMO, Code_Saturne, Cybele) **[M12] [BSC, STFC, CYB]**

D5.2(a,b,c) Implementation of required modifications to each application (XIOS/NEMO, Code_Saturne, Cybele) **[M24] [BSC, STFC, CYB]**

D5.3 In-depth analysis of I/O characteristics of UEABS benchmark suite and Skeleton benchmarks suite **[M24] [STFC]**

D5.4(a,b,c) Optimized version of each application (XIOS/NEMO, Code_Saturne, Cybele) **[M36] [BSC, STFC, CYB]**

WP6	Lead beneficiary			**ICCS**			
Work package title	Acceleration, integration, and evaluation of the converged scalable architecture						
Participant number	1	2	3	4	5	6	7
Short name of participant	FORTH	BSC	STFC	BULL	JGU	**ICCS**	CYB
Person/months per participant	9	12	20	28	7	43	8
Start month	1		End month		36		

The main objectives of this WP are the followings:

Objectives

- Continuous integration of software and hardware components in a single stack for performance evaluation.
- Development of the rack architecture for the converged I/O path with in-transit acceleration.
- Development of a high-level architectural simulator to explore I/O behavior at large scale.
- Design and implementation of hardware accelerators that will be used for in-transit data transformations.
- DIO technology evaluation.

Description of work

Task 6.1 Prototype for continuous integration and evaluation of system components [M1-M36] [BSC, All Partners**]**

This task will create a skeleton for the prototype stack where individual components and applications can be integrated continuously for evaluation and optimization purposes. This framework will initially include mock-up components. As project technology is implemented via the different components partners will be able to use the prototype skeleton to replace previous version of their components and perform further evaluation and optimizations. This skeleton will also be important for identifying early on gaps in the design and implementation issues that need to be dealt early on for timely delivery of technology.

Task 6.2: High-level fully functional architectural simulator [M1-M18] [STFC, ICCS, BULL**]**

In this task we will extend the Prazor[104] high-level, full-system simulator used by STFC, with support to simulate

[104] DJ Greaves, M Puzovic: Prazor/VHLS User Manual. https://goo.gl/svq6Fj

I/O. This simulator will use novel techniques that make micro-architectural and peripheral design-space exploration of performance, power consumption and energy use, feasible and accurate at different scales. It is compatible with the real hardware meaning that it can run the same operating system and use the same disk images as the real hardware that is simulating. Moreover, the target hardware (micro-architecture and peripherals) is built using highly modular and extensible virtual prototyping blocks (VPB). To accomplish this task we will implement additional VPBs in the simulator that will be used to simulate novel I/O peripherals. For example, we will add logic to contemporary I/O peripherals that will be able to do near-data computation such that necessary filtering and mapping functions are performed where the data is stored instead of moving it to processor to operate on. We will do this work early on so that other parts of this work package can use our I/O extensions for rapid prototyping and to understand the trade offs, e.g. before implementing the final acceleration solution on FPGAs. Throughout the rest of the time we will work on this task in parallel with other tasks by adding further features to I/O peripherals and functionally testing the software stack before being applied to the real system. We will conclude this task by calibrating all newly added VPBs and developing theory how to scale them such that they reflect future performance- and energy-critical characteristics of exascale platforms so that we can use the simulator in T6.6 to extrapolate results from this project to the future exascale machines.

Task 6.3 Design space exploration of data transformations with respect to accelerator, storage, and processing alternatives [M1-M12] [ICCS, BULL**]**

In this task, we will measure and analyze the basic costs in a system with respect to the relative location of the accelerator, storage, memory, and processing, based on technology projections. We will examine the overhead, latency, throughput, *energy*, and complexity of each alternative. We will also take into account, the storage abstractions and transformations developed in WP2 and the storage abstractions and transformations developed in WP3 to identify the main parameters with respect to HPC applications, reduce storage overhead and improve acceleration efficiency. We will study how data layout affects basic and application costs and we will identify popular types of transformations that can significantly affect efficiency of future systems. To identify the Pareto optimum architecture based on the application requirements, we will perform a design space exploration taking into account traffic requirements and architectural constraints.

Task 6.4 FPGA accelerator for cut-through data processing and layout transformations [M7-M24] [ICCS, BULL**]**

Based on the design spaces exploration in T6.3, this task will focus on the implementation of the accelerators for the DIO I/O stack. We will implement the hardware accelerators (kernels) for the FPGA boards required for on-the-fly data processing as per the analysis and requirements of WP3 with respect to application needs, including data transformations in layout and content. Furthermore, we will explore hardware accelerators for certain operations of the key-value store itself that have the potential to offload overhead from the host processors and to increase efficiency, including predicate selection and certain types of comparisons that exhibit parallelism. Finally, we will also develop the required interfaces, controllers, and glue software at the host level so that the DIO software stack can seamlessly initialize, use, and dynamically manage the FPGA kernels. In this task we will explore both FPGA accelerators that are based on PCI (popular today) as well as FPGAs coupled with the host CPU, an emerging technology that is starting to be available[105]. Finally, this task in particular will examine issues related to data transfers among host memory, accelerators, and storage devices to optimize related costs given technology and architectural constraints.

Task 6.5 Evaluation of integrated prototype including I/O and acceleration and the software stack [M25-M36] [ICCS, All Partners**]**

In this task, we will use the DIO software stack prototype and the integrated accelerators to evaluate the converged rack-scale system. Several partners have available racks of servers with fast devices, where we will use the systems software stack, fast storage devices, and accelerators and we will evaluate in depth all aspects of the proposed technology. The software stack will include the full I/O path, applications, APIs and libraries, and resource management developed. Our evaluation will cover high level operations at the application level, e.g. create, load datasets, management operations, e.g. for load balancing, concurrent applications in mixed workloads that need shared access to system resources, transparent acceleration of tasks for the applications in the project, and checkpointing. Furthermore, during the evaluation, we will examine several technology aspects of the work by analyzing tradeoffs and projecting to future technology trends.

Task 6.6: Extrapolation to large scale systems [M25-M36][STFC, All Partners**]**

[105] Intel Hardware Accelerator Research Program – A Tutorial for learning and using the Intel Xeon with integrated FPGA. Elizabeth Barnes, et al. ISCA 2017. June 24-28, 2017, Toronto, ON, Canada.

In this task we will use the simulator from T6.2 and measurements in the project from the actual prototype to extrapolate performance to larger scale systems. We will also extrapolate the skeleton benchmarks and real-application performance evaluation results to exascale. To accomplish, this we will extend high-level architectural simulator from T6.2 to accept traces from large-scale production-sized executions and also scale-up *virtual building blocks* of a simulator in order to be able to model performance- and energy- critical characteristics of the future exascale platforms. The focus in this evaluation will be on the data movement and I/O rather than processing. We will evaluate in detail various systems aspects, including: cache mechanism, policies, data placement, concurrency, isolation, replication, libraries overhead, and checkpointing.

Deliverables

D6.1 Prototype skeleton for continuous software integration and evaluation of system components **[M12] [BSC]**
D6.2 Extension of high-level fully functional architectural simulator with I/O simulation. **[M18] [STFC]**
D6.3 Design space exploration for accelerated data transformations **[M12] [ICCS]**
D6.4 Implementation of hardware accelerators for on-the-fly processing **[M24] [ICCS, BULL]**
D6.5 Evaluation of integrated prototype and extrapolation to larger scales **[M36] [ICCS, STFC]**

WP7	Lead beneficiary			**BULL**			
Work package title	Dissemination and Exploitation						
Participant number	1	2	3	4	5	6	7
Short name of participant	FORTH	BSC	STFC	**BULL**	JGU	ICCS	CYB
Person/months per participant	7	4	4	10	4	4	6
Start month	1		End month		36		

Objectives

This WP includes and coordinates all dissemination, exploitation, and communication activities to be performed by the consortium. The goal is to ensure broad dissemination of project results, activities, and accomplishments to different audiences and to form a focal point for exploitation activities. Dissemination includes both technical audiences for research results and achievements and communication to the general public for awareness on technology related to HPC in general and storage in particular. Overall, the objectives of the WP are:

- To prepare the dissemination and communication strategy and plan for the project.
- To prepare communication material for the project: Web page, social media presence, presentations, leaflets, punch-line messages, audio-visual materia, short demos and teasers, etc.
- To oversee technical publications for different audiences and in appropriate venues, e.g. international conferences for research purposes and industrial press for commercial awareness.
- To raise awareness towards the key actors in the field and other appropriate audiences through appropriate communication.
- To ensure communication (and training) towards applications communities and infrastructure designers, ensuring coverage of the end-to-end value chain.
- To prepare and maintain a Data Management Plan for issues related to the data that will be used or generated by the project.
- To prepare and maintain an exploitation plan for the project technology and pursue its implementation.
- To examine opportunities for standardization activities and monitor their progress.

Description of work

This WP measures the project exploitation and the success of the DIO platform and services. All technology and R&D partners will participate in the dissemination and exploitation activities of the WP.

Task 7.1 Dissemination and Communication Strategy and Plan and Data Management Plan [M1-M3, M12, M24, M36] [BULL, All Partners**]**

This task will prepare early in the project (by M3) dissemination and communication strategy and plan and will update this document periodically (M12, M24, M36) based on developments and outlook in the project. The purpose of the strategy and plan is to define the project communication messages, channels, and audiences with respect to:

- Research and technical activities
- Technology roadmaps in industry and academia
- Domain specific application communities

- Awareness on HPC-Cloud convergence
- Monitoring coverage of the end-to-end value chain
- Training and education
- Public awareness
- Standards

In building the DIO community the plan will use the collaboration and research networks of individual partners, which will ensure EU-wide coverage. Consortium partners have a strong reputation in their respective areas of activity which ensures broad access to the research community in Europe as well as to European and global industry. In addition, each partner will be responsible for general public and education activities in their geographic area, ensuring breadth. As part of this task we will also compile the Data Management Plan (DMP) of the project. This plan will evolve during the lifetime of the project, as work and requirements progress.

Task 7.2 Dissemination and Communication Material [M1-M6] [FORTH, ALL Partners**]**

This task will produce the material required for dissemination and communication actions, and which will be refined during the project communication actions. To simplify the dissemination and communication process and provide a uniform visual profile to the world, we will develop stationary material (leaflets, templates, poster) for different events (e.g. conferences, workshops, and exhibitions, discussions with industry, etc.). The material includes:

- Project web page
- Project logo and other visual material
- Project initial leaflet, poster, and templates
- Social media presence setup
- Project Tools and process for project audio-visual material

Task 7.3. Dissemination and Communication Actions [M1-M36] [BULL, ALL Partners**]**

This task includes all actions to ensure broad communication to the identified audiences, project visibility at different levels, and to generate traction of project technology in different communities. The actions to perform during this task include:

- **Continuously update the web page with material on project technology, results, and activities.**
- **Periodic presence in social media with short updates on project activities and results.**
- **Periodic updates to an opt-in mailing list with updates on project activities and results.**
- **Occasional press releases for project achievements.**
- **Publications, participation, and presentations in conferences, workshops with technical content.**
- **Articles for the technical and general-public press for awareness purposes, including national press.**
- **Interaction with standards groups for I/O libraries (NetCDF, HDF5, XIOS) and SLURM.**
- **Demos for project technology at different levels (more or less technical)** for conferences and technical venues, but also for the web targeting general public awareness. Demos will use different media, including audio-visual material.
- **Synergies with existing initiatives and projects.** We will use partner collaboration networks to facilitate synergies with other activities and projects including: EU projects, concertation events, contributions to vision meetings and roadmaps, contributions to position papers, analysis of trends, participation in working group meetings for prioritization of challenges and technology updates. Project partners have a strong presence in the areas of storage systems and I/O and therefore, can perform extensive actions in this direction.
- **Standardization.** This task will also cover the progress of standardization activities during the project and any follow-up activities planned or continuing after project completion.
- **Industry day.** The project will organize an industry day using STFC's approach[97] to bring technology and needs together. The industry day will focus on DIO technology.

Task 7.4: Management of IPR, Exploitation and Commercialization [M7-M36], [BULL, All Partners**]**

This task will coordinate the exploitation activities of the project. First, this task will examine and clarify IP issues for any future use of the DIO prototype or specific components and their interactions. This is important especially in DIO where the proposed technology centers around systems software in a complex ecosystem of HPC layers and components.

This task will draft a roadmap with projections from project findings. One of the goals of the project is to provide visibility to the capabilities of optimized I/O stacks for future systems. For instance, to provide realistic answers to questions such as: how much should the overhead be per I/O operation? What is the cost of moving data in different cases? How fast can exploration of datasets become in project applications?

The task will also draft a plan for follow-up Proof-of-Concepts (PoCs) as discussed in Section 2 (Impact). The purpose of these PoCs will be to further validate technology after the end of the project in operational environments and to understand remaining gaps.

Deliverables
D7.1 Dissemination and communication strategy and plan **[M03] [BULL]**
D7.2(a,b,c,d) Data management plan (and revisions) **[M03, M12, M24, M36] [FORTH]**
D7.3 Project web page, project stationary, and other dissemination and communication tools **[M06] [FORTH]**
D7.4(a,b,c) Dissemination and communication activities **[M12, M24, M36] [BULL]**
D7.5(a,b,c) Management of IPR, Exploitation and Commercialization plans, including roadmap and PoCs **[M12, M24, M36] [CYB]**
D7.6 Industry day to present project technology and plans **[M36] [STFC]**

Table 3.1b: List of work packages

WP No	WP Title	Lead Participant No	Lead Participant Short Name	Person-Months	Start Month	End month
1	Project Management	1	FORTH	26	1	36
2	Data storage and access mechanisms	1	FORTH	92	1	36
3	Storage abstractions and I/O libraries	5	JGU	101	1	36
4	Resource and QoS management	2	BSC	84	1	36
5	Application adaptation and evolution	3	STFC	84	1	36
6	Acceleration, integration, and evaluation of the converged scalable architecture	6	ICCS	127	1	36
7	Dissemination and exploitation	4	BULL	39	1	36
			Total PMs	553		

Table 3.1c: List of Deliverables [R: Report, O: Other (e.g. software), DEC: website, press/media actions].

Deliverable number and name		WP No	Lead participant short name	Type	Dissemination level	Delivery month
D1.1	Public project presentation	WP1	FORTH	R	PU	M1
D1.2	Quality management plan	WP1	FORTH	R	PU	M3
D1.3	First periodic report	WP1	FORTH	R	CO	M12
D1.4	Second periodic report	WP1	FORTH	R	CO	M24
D1.5	Third periodic report	WP1	FORTH	R	CO	M36
D1.6	Final project report	WP1	FORTH	R	CO	M36
D2.1	Design and implementation of the memory mapped key-value store	WP2	FORTH	R+O	PU	M12
D2.2	Design and implementation of the scale-out key-value store	WP2	FORTH	R+O	PU	M24
D2.3	RDMA-based replication and data transfer mechanisms for data reorganization	WP2	STFC	R+O	PU	M24
D2.4	Final prototype of optimized key-value store	WP2	FORTH	R+O	PU	M36
D3.1	Requirements of applications and standard I/O libraries over DIO	WP3	JGU	R	PU	M6
D3.2	Design of DIO multi-level checkpointing library for heterogeneous architectures	WP3	BULL	R	PU	M6

D3.3	First prototype of DIO multi-level checkpointing library	WP3	FORTH	R+O	PU	M12
D3.4	First adaptation of standard I/O libraries over DIO	WP3	JGU	R+O	PU	M12
D3.5	Full implementation of DIO multi-level checkpointing library	WP3	FORTH	R+O	PU	M24
D3.6	Novel APIs for data transfer and transformation in systems with deep storage hierarchies	WP3	JGU	R+O	PU	M24
D3.7	Final adaptation of standard I/O libraries and novel APIs over DIO	WP3	JGU	R+O	PU	M36
D4.1	Design and implementation of non-intrusive tool for I/O and accelerator profiling	WP4	STFC	R+O	PU	M12
D4.2	Design and preliminary implementation of storage scheduler	WP4	BSC	R+O	PU	M12
D4.3	Complete implementation of the storage scheduler for dynamic resource allocation under high concurrency	WP4	BSC	R+O	PU	M24
D4.4	Design and implementation of scheduler-based data reorganization techniques	WP4	BSC	R+O	PU	M36
D5.1[a,b,c]	Design of required adaption for each application (XIOS/NEMO, Code_Saturne, Cybele)	WP5	BSC, STFC, CYB	R	CO	M12
D5.2[a,b,c]	Implementation of required modifications to each application (XIOS/NEMO, Code_Saturne, Cybele)	WP5	BSC, STFC, CYB	R+O	CO	M24
D5.3	In-depth analysis of I/O characteristics of UEABS benchmark suite and Skeleton benchmarks suite	WP5	STFC	R	PU	M24
D5.4[a,b,c]	Optimized version of each application (XIOS/NEMO, Code_Saturne, Cybele)	WP5	BSC, STFC, CYB	R+O	CO	M36
D6.1	Prototype skeleton for continuous software integration and evaluation of system components	WP6	BSC	R+O	PU	M12
D6.2	Extension of high-level fully functional architectural simulator with I/O simulation	WP6	STFC	R+O	PU	M18
D6.3	Design space exploration for accelerated data transformations	WP6	ICCS	R	PU	M12
D6.4	Implementation of hardware accelerators for on-the-fly processing	WP6	ICCS, BULL	R+O	PU	M24
D6.5	Evaluation of integrated prototype and extrapolation to larger scales	WP6	ICCS, STFC	R	CO	M36
D7.1	Dissemination and communication strategy and plan	WP7	BULL	R	CO	M3
D7.2[a,b,c, d]	Data management plan (and revisions)	WP7	FORTH	R	PU	M3, M12, M24, M36
D7.3	Project web page, project stationary, and other dissemination and communication tools	WP7	FORTH	DEC	PU	M6
D7.4[a,b,c]	Dissemination and communication activities	WP7	BULL	R+ DEC	PU	M12, M24, M36
D7.5[a,b,c]	Management of IPR, Exploitation and Commercialization plans, including roadmap and PoCs	WP7	CYB	R+ DEC	CO	M12, M24, M36
D7.6	Industry day to present project technology and plans	WP7	STFC	R+	RE	M36

3.2 Management structure, milestones and procedures

3.2.1 Management Principles

Project management is designed around three principles:

(i) Simple structure to allow quick and timely decisions;
(ii) Well-defined mechanisms to resolve problems and conflicts in a fair manner; and
(iii) Clear responsibilities for each actor for accountability and risk management.

Partners will use standard management techniques i.e., PERT charts, periodic internal progress reports, and milestone checking, to monitor the project progress against the proposal.

3.2.2 Organisational structure

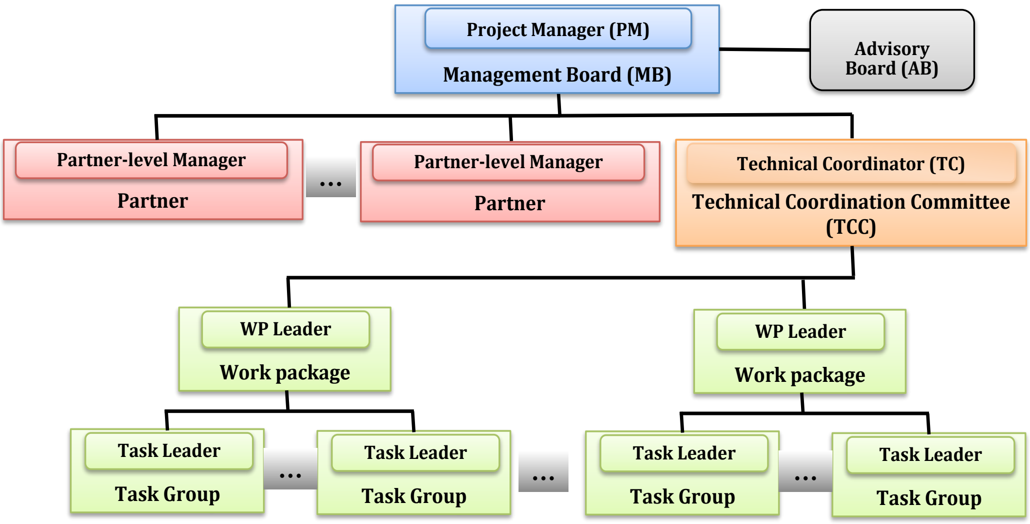
The project has three relevant structures:

Technical Coordination Committee (TCC): The technical coordination structure of the project (Figure 15) has, in a bottom-up fashion, the following layers:

- Small task-groups have the goal of tackling individual, well-defined tasks. Each task-group consists of a number of researchers and engineers that will work on the specific task. Members of each task-group may (and usually will) originate from different partners. Each task-group has a leader that interfaces with other leaders in a peer-to-peer manner and with higher levels in the project structure.
- Each WP consists of a set of tasks. Each WP has a leader. The WP leader is responsible for interfacing to other WP leaders in a peer-to-peer manner and to higher levels in the project structure.
- The project has a single technical coordinator, who originates from the coordinating partner (FORTH). The (technical) project coordinator is chosen based on his technical background and experience from leading technical teams.

The technical structure includes two bodies:

- Technical coordination committee (TCC). The project TCC consists of the technical coordinator and the WP leaders. The TCC meets approximately once every three months to discuss the current status and validate or revise current planning.



3.2.3 Figure 15: Technical coordination and project management structure.

- Technical general assembly (TGA). The project TGA consists of all technical personnel, researchers and engineers that work in the project at any point in time. We expect that the project TGA will only meet a few times during the project, and at most once per year, if necessary.

Task leaders, WP leaders, and the technical project coordinator are appointed by the project management board, as discussed next.

Management Board (MB): The management structure of the project (Figure 15) has, in a bottom-up, fashion the following layers:

- Each partner has a partner-level manager, appointed by each partner. Partner-level project managers are chosen mainly based on their experience with large teams in research and product groups and secondarily based on their technical background.
- The project has a single project manager, who is the partner-level manager of the coordinator (FORTH).

The management structure includes one body, the management board (MB). The project MB consists of the individual partner project managers (and the project manager) and the project technical coordinator. Please note that the project technical and management structures and the bodies they encompass are being assembled with well-defined procedures, are known upfront, and do not have any uncertainties. Also, given this structure, the coordinating partner effectively undertakes the full technical, financial, and administrative management responsibility for the project, providing a single contact point to the EC and external actors.

Advisory board (AB): Finally, the project management structure includes an Advisory Board with the purpose of providing feedback about project goals and progress to the management board. Additionally, the Advisory Board will also function as a link in the end-to-end value chain to provide bridges with important actors in industry and academia and to assist with awareness related to project technologies. Our goal is to use a "lightweight" advisory board mechanism with two or three members, focusing as much as possible on technology and its impact. The members of the board will originate from applications and storage, in an effort to tie together the two ends of the software stack, with experience in technical management. The project budget foresees an amount in the order of 5000 Euros for expenses for the advisory board, as part of the coordinators travel budget.

3.2.4 Roles

Technical roles are assigned in a well-defined manner in each WP description. These roles form a closed-loop system as follows:

- The project TCC is responsible for the high-level project design and directions.
- Each WP leader and all task leaders within a WP are responsible for drafting the WP-level design, following the general design of the TCC.
- Each task leader and task members are responsible for drafting task implementations, based on WP design.
- Individual task members' work based on the task implementation plan.
- Task leaders supervise work within each task and report progress, problems, and suggestions to the WP leader.
- WP leaders supervise work within each WP and report progress, problems, and suggestions to the TCC. **Quality assurance (QA)** for internal interfaces and deliverables between tasks and WPs are the task of WP leaders. The leader of the WP that will use an interface produced by another WP will assign internal reviewers. Internal interfaces and deliverables will have a QA-deadline of one week before delivery.
- The project TCC is responsible for high-level design decisions for the full project, reviewing technical progress, assuring quality of deliverables and technical work in general, examining technical problems that arise, and hearing suggestions made by technical staff. Finally, the TCC, through the technical project coordinator is responsible for raising to the management board issues that may affect the project implementation and require intervention of the management board.

  In particular, **quality assurance (QA)** for external deliverables is the role of the TCC**.** Deliverables produced by work package leaders will be submitted to the TCC for quality assurance. The TCC will assign internal reviewers that will check the quality of the each piece of technical work. External deliverables will require a QA-deadline of two weeks before delivery.
- The *Advisory Board (AB)* provides assistance with respect to technology, trends, and potential to the Management Board.

Management roles of each entity in the management structure are as follows:

- The management board is responsible for ensuring the proper execution of the project within the time and financial resources provided to the project. Thus, the management board is responsible for taking decisions related to issues, such as:
  - Intellectual property and innovation management
  - Reviewing technical progress as presented by the TCC
  - Ensuring timeliness of deliverables and results
  - Hearing and managing changes to project goals, as suggested by the TCC
  - Interacting with the Advisory Board, with the EC, and external actors
  - Managing modifications to the project consortium

  The Management Board will meet at least three times a year. Additional meetings may be arranged based on project needs. Meeting may collocate with other events for efficiency purposes.
- The project manager is responsible for interfacing to external entities and actors for reporting, dissemination, and exploitation purposes. The project manager may assign specific tasks to other project members, but bears the responsibility of all these tasks, as far as the project is concerned.

3.2.5 Meetings

Provision will be made in the planning for at least the following types of meeting:

(i) MB, TCC, and AB meetings (at least three times a year, possibly joined);

(ii) TGA workshops (once per year);

(iii) EC Review Meetings (once per year);
(iv) Working visits of partners (on a need-basis)

Besides physical meetings, partners will also use teleconferencing/videoconferencing facilities for discussions and reporting. All partners are familiar with such procedures and are able to use them judiciously for best efficiency.

3.2.6 Decision making mechanisms

The project structure provides all partners with the opportunity to participate in all decision-making processes (both technical and administrative). This will hopefully lead to decisions based on discussion and consensus at each level. However, in cases where this is not possible, the structure enforces:
(a) Hierarchical propagation of conflicts that are not resolved by consensus to a single entity (MB)
(b) Conflict resolution using a voting mechanism in the MB.

Rach issue that arises in technical or administrative tasks and is not resolved by consensus is propagated to the next level, along with a set of alternative solutions/approaches. If a technical issue reaches the TCC and is not resolved it is propagated to the MB. At the MB all issues are resolved either by consensus or voting, based on the alternatives presented.

Therefore, decisions in both technical and management bodies will be resolved with a simple majority voting scheme. Each member in each body has a single vote (there are 7 partners in DIO). We should note that every effort will be made to reach decisions with consensus based schemes. We believe, based on previous experience that this is in most cases possible. However, should the need arise, decisions will be taken based on a vote. The Consortium Agreement will also include all necessary provisions.

3.2.7 Appropriateness for the project

The structure of the consortium is based on previous experiences about technical coordination and project management in both research and industrial projects of the size of the consortium. First, we should note that the consortium has a manageable size and would correspond to a mid-size team in an industry project and a mid-to-large team in an academic project. Given this, it makes sense to have a tight interaction model, with few levels in the technical and management structures.

In fact, the size of the consortium was chosen with this in mind. We believe that critical technologies need to be developed by medium sized teams that can quickly coordinate among themselves, review progress and results, and define the next steps in the design and implementation processes. The proposed structure serves exactly this goal. Each task is a well-defined piece of work that can be tackled in a short amount of time by a small group (2-4 people). Results are quickly communicated at the next level up, the WP, and are incorporated in the rest of the WP tasks. Each WP at any point in time consists of a small number of tasks (1-4). Finally, WP progress is discussed at somewhat longer time intervals, which is appropriate for the looser interactions between WPs.

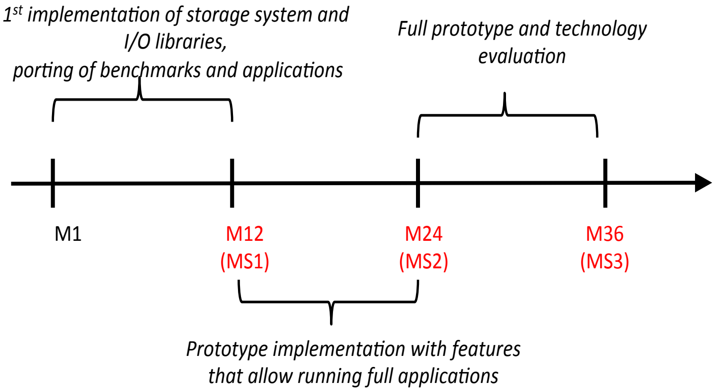

1st implementation of storage system and I/O libraries, porting of benchmarks and applications
Full prototype and technology evaluation
M1
M12 (MS1)
M24 (MS2)
M36 (MS3)
Prototype implementation with features that allow running full applications

Figure 16: DIO at a glance - Duration and high-level milestones.

The management structure of the project is thinner, to allow (1) each partner to have a complete view of the project at any point in time, (2) to facilitate quick reaction when any issue is brought up by a partner, and (3) each body and role to have a clear goal and interface with the other bodies and roles. A single body, the management board, oversees progress and, as there is no other layer in the management hierarchy, is able to directly see and react to any problem or deviation.

Overall, we expect that the proposed technical coordinator and management structures will be very effective, with the assistance of the Advisory Board, in dealing with issues that may arise in the duration of the project.

3.2.8 Milestones

Figure 16 shows a high-level view of the project time-line. Table 3.2a includes a list of the main milestones that will be used to structure the work of the project at a high level and to monitor progress.

Table 3.2a: List of milestones.

Milestone number and name		Related WPs	Due (month)	Means of verification
MS1	First implementation of storage system, I/O libraries, application, benchmark porting	WP2,3,4,5,6	M12	Sample runs of first prototype

MS2	Prototype implementation with features that allow running full applications	WP2,3,4,5,6	M24	Sample runs of real applications
MS3	Full prototype and technology evaluation	WP2,3,4,5,6	M36	Technology evaluation with real applications

RISK MANAGEMENT AND CONTINGENCY PLANS

The project will implement a risk management plan based on a rigorous and continuous risk analysis methodology involving all consortium members. The Management Board will assess global risks based on the information delivered by the WP Leaders and the Technical Coordinator. WP Leaders will take the responsibility to identify and report risks, which threaten the achievement of WP objectives within the planned time and financial budget. Furthermore, WP Leaders will prepare periodic analysis of possible consequences for the project future achievements and develop a proposal for managing the risk, to be discussed in an emergency Management Board meeting. Based on the results of the analysis and of additional factors such as the probability of the identified risk to occur and the importance of its prospective impact, the Management Board will provide the consortium, in order of priority, with one or more of the following plans, which shall be implemented by the partners:

- Avoidance plan proposing solutions to prevent the anticipated problems.
- Mitigation plan with workarounds to lower the impact on the foreseen technological and scientific objectives.
- Contingency plans with strategies on how to minimise the risk impact once it has occurred.

Therefore, prior to beginning a substantial technical activity, a documented risk analysis will be carried out by the WP Leader in charge. This analysis will focus on time allowed, cost, functionality, quality, and use of resources.

Table 3.2b presents the main possible global risks for the project together with probability evaluation, assumed impact on the project progress/achievements, and contingency strategy to avoid the occurrence of the anticipated problems. The table will be continuously updated during the entire life-time of the project.

Table 3.2b: Critical risks for implementation.

Description of risk	Likely-hood: Low/Med/ High	WP(s) involved	Impact	Proposed risk-mitigation measures
Development of system components, such as key-value store, requires more effort than planned	Medium	WP2,3,4,5,6	Provide fewer optimizations	Partners are experienced in this type of work and have resources to prototype the required components. In addition, DIO applies continuous integration to provide a prototype in multiple stages with increasing functionality. Therefore, DIO will prototype the technology and evaluate it, even if not all planned features are included.
Integration with I/O Libraries is more complex than expected	Medium	WP3	Reduced impact due to more limited scope of work on I/O Libraries	DIO uses an approach of continuous integration, starting from a "mockup" stack that includes all components. This approach reveal early any gaps in I/O library implementation. If not possible to cover gaps with shifting resources, partners will limit work on I/O libraries to a subset covering project applications.
Integration with other system-level components, requires more time than estimated	Low	WP2,4	More time to complete project	Incremental staging of modifications to allow for clear and measurable estimation of improvements throughput project execution
Key-value store approach for I/O has limited benefits				
Complications in implementing multi-level checkpointing with heterogeneous accelerators	Medium	WP3,5	More time to complete corresponding tasks	Proactive study of available programming tools and interfaces to save/restore accelerator state, drawing on the available extensive experience within the consortium in the internals of checkpointing libraries and in the utilization of FPGA-based accelerators.

Applications are not able to stress the system to the levels required as is	Low	WP6	Will not be able to generate the required load	Applications under consideration have a wide range of parameters. Partners will identify the right parameter range to match future storage system needs.
Storage devices required for demonstrating technology benefits not available or problematic	Low	WP6	Target performance will need to be scaled down to existing devices	Partners will use NVMe cards available at the time and mix with multiple types of SSDs (single cell and multi cell devices that exhibit different performance characteristics) to produce a deep storage hierarchy with devices existing at the time of proposal execution
FPGA integration with server processors not available during the project	Low	WP6	Evaluation will be limited to FPGAs that are plugged in the I/O path and exist today	Extrapolate numbers from I/O-attached accelerators to processor-attached accelerators via detailed breakdowns
Concurrent and more competitive technologies are developed elsewhere	Low	WP7	Temporal loss of leadership and change of exploitation conditions	Monitor the evolution of the field in both the scientific and product areas. Development of a modular system so that new technologies can easily be integrated in different system levels
Industrial partner is taken over by another company	Low	WP1	Delay and even modification of exploitation plans	Definition of precise rules in the consortium agreement for modified roles or replacement of partners
A partner leaves the consortium	Low	WP1	Delay in partner contribution and integration of its results	Monitor the progress of each partner in the project. Search for a partner substitute with the right expertise or taking over the partner responsibilities by other consortium partners

3.3 Consortium as a whole

3.3.1 Partner expertise coverage

The consortium consists of five research academic organizations (FORTH, ICCS, JGU, STFC, BSC) out of which two are major HPC centers (STFC, BSC) and JGU operates a Tier-2 facility as well for Germany, and two industrial partners (BULL, CYB), from five EU countries. The following table summarizes the partner geographic precedence and expertise.

Participant number and short name		Country	Expertise
1.	FORTH	Greece	Systems software, storage systems, storage architectures, RDMA networks, HPC runtime systems
2.	BSC	Spain	Parallel file systems, storage systems, scheduling, parallel I/O, HPC storage, HPC applications (NEMO group), I/O libraries, load balancing, data placement. BSC also operates a Tier-0 HPC facility.
3.	STFC	UK	HPC architectures, HPC hardware/software co-design, HPC runtime systems, HPC applications. STFC operates a Tier-1 facility.
4.	BULL	France	HPC systems and architectures
5.	JGU	Germany	Fast storage devices, persistent memory, Parallel I/O, Parallel filesystems, I/O libraries, load balancing, data placement. JGU also operates a Tier-2 facility.
6.	ICCS	Greece	FPGA accelerators, customized memory hierarchies, data layout transformations, memory architectures, data transfers
7.	CYB	France	HPC applications in agricultural production forecast and plant modeling, Numerical methods, Machine learning

The key criterion for the specific selection of partners for the consortium has been their outstanding skills in research and/or technology development in the areas of I/O, storage systems, storage device technology, HPC architectures, and HPC applications, and operational aspects. Their expertise balances among three important poles: *technology, system design and implementation,* and *market needs and trends.*

- **Technology:** Designing and implementing scalable high-performance I/O systems in a cost-effective manner requires understanding of how technology will evolve in the next years and the reasons behind technology trends. The partners of the project, due to their experience in the areas of I/O, storage systems, storage device technology, HPC architectures, and HPC applications, and operational aspects and their previous and current work are in unique position to understand issues and to make projections about trends. All partners are involved already in numerous activities related to storage technologies, HPC exascale architectures, applications, and performance, and have a clear view of technology trends in their respective fields, including the system and the application level.

- **System design and implementation:** Besides understanding technology trends in this area, it is important to also have a demonstrated capability to design and implement solutions. This is extremely important in storage systems as they involve low-level, critical system software that is very challenging in its nature. All partners in the project have demonstrated this capacity in numerous previous projects, and in fact, have formed this consortium based on this criterion.
- **Market needs and trends:** Besides the ability to project technology and to design and demonstrate solutions, it is important to be able to understand market needs and trends. This is of paramount importance when evaluating the timeliness of a solution: A solution, no matter how technologically advanced, is bound to fail if it is not timely. For this reason the consortium includes BULL and CYB that are currently active in commercial HPC systems and applications, respectively. In addition, BSC and STFC are both major users of HPC systems for data intensive applications. More importantly, BULL, STFC, and BSC are involved in designing/procuring next generation systems and are fully aware of the most pressing issues in achieving the next-generation performance for I/O. Finally, FORTH has been in direct contact with market and business needs over the last six years by commercializing multiple rounds of research results and delivering new products in collaboration with Industry. Thus, all project partners are well in-line with market needs and market trends and the steps and effort required to take research results to market.

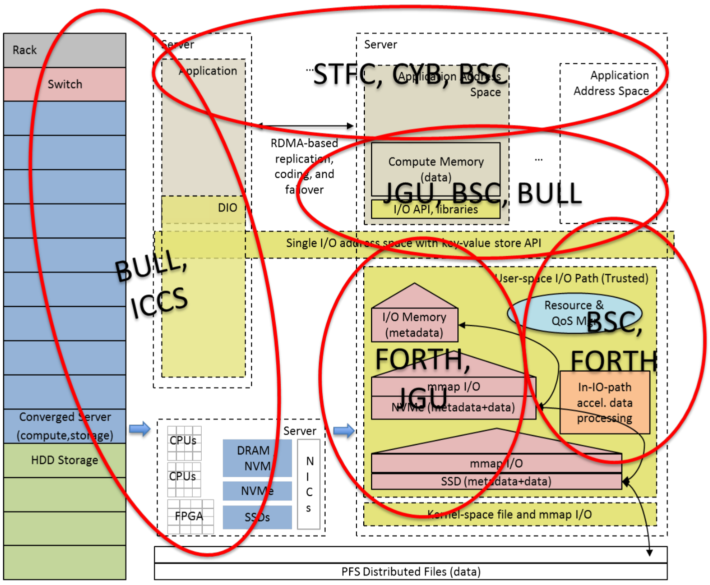

Rack
Switch
Converged Server (compute,storage)
HDD Storage
Server
Application
DIO
RDMA-based replication, coding, and failover
Server
Application Address Space
Application Address Space
Compute Memory (data)
I/O API, libraries
...
Single I/O address space with key-value store API
User-space I/O Path (Trusted)
I/O Memory (metadata)
Resource & QoS Mgt
mmap I/O
NVMe (metadata+data)
In-IO-path accel. data processing
mmap I/O
SSD (metadata+data)
Kernel-space file and mmap I/O
Server
CPUs
CPUs
FPGA
DRAM NVM
NVMe
SSDs
NICs
PFS Distributed Files (data)
STFC, CYB, BSC
JGU, BSC, BULL
BULL, ICCS
FORTH, JGU
BSC, FORTH

Figure 17: Mapping of partner expertise to components.

3.3.2 Project coordination

FORTH has had extensive expertise both as a coordinator and as partner in numerous FP7 and H2020 projects in the recent years. In the area of computing systems specifically, the CARV Laboratory of FORTH-ICS was the project and technical coordinator for IOLANES (2010-2013), the project coordinator for ASAP (2014-2017), and is currently coordinating ExaNest (2015-). Furthermore, CARV (within FORTH-ICS) is coordinating numerous major technical activities within EC projects where it has recently participated (or is participating): STREAMS, CumuloNimbo, NanoStreams, CoherenPaaS, LeanBigData, EuroServer, Vineyard, ExaNode, BigStorage, and several national projects. The coordinating researcher on behalf of FORTH, *Prof. Angelos Bilas,* has extensive prior expertise in coordinating sizeable research teams acquired during previous research projects abroad and in Europe and has participated to numerous Marie Curie actions (including an excellence grant and via coordinating the very successful IOLANES projects (2010-2013) that achieved high impact by commercializing its technology. He is currently managing a team of about 20 graduate students, postdoctoral fellows, and staff members. Finally, FORTH has an efficient administrative structure that will support the project in all administrative tasks.

3.3.3 Task assignment to partners and complementarity/coverage of expertise

The project combines three types of partners:

- Industrial: BULL and CYB
- HPC centers: BSC and STFC
- Academic: FORTH, JGU, and ICCS

The academic partners FORTH, JGU, and ICCS, but also the Supercomputing Centers, BSC and STFC, bring their research experience on systems software, storage systems, parallel I/O, FPGA accelerators, emerging storage devices including FLASH and NVM, storage architectures, parallel applications, HPC I/O, and interconnection networks. The industrial partners, BULL and CYB but also HPC centers STFC and BSC contribute their industrial and operational experience in the areas of HPC infrastructures, HPC storage architectures, storage devices, I/O path systems software, and HPC applications. Partners complement each other very well, covering all required expertise, as also summarized in Figure 17.

3.4 Resources to be committed

The resources committed to the project by each partner fall mainly in three categories: personnel (technical and management), travel, and equipment, with small costs for other goods and services. Next, we discuss resource allocation to WPs and partners.

3.4.1 Overview

The DIO budget and requested EC funding amounts to about 3.99M€. Most resources are allocated to personnel costs (91%), followed by travel (4.3%), and equipment (3.8%). A small budget percentage has been allocated to other goods and services (0.9%). Next, we discuss each category.

3.4.2 Technical and management staff effort

The person effort required to build the technology proposed in this project is estimated to about 45 person years spread over a period of three years. Out of these 45 person years, 88% is allocated to R&D tasks, 5% is allocated to management, and 7% to dissemination and exploitation activities. The effort required is divided in terms of seniority as follows: We expect about one quarter to be senior engineers and researchers, whereas the other three quarters to be divided between mid- and junior-level researchers and engineers. Senior personnel will be mostly involved in design issues and will tackle the implementation of challenging system components. Mid-level personnel will mostly deal with implementation of the system and revision of system designs. Finally, junior personnel will carry out implementation tasks, guided by mid-range and senior personnel.

This outlined technical person effort is adequate to carry the project. We should note that building low-level systems software that is in the critical performance path is challenging in many respects and may require significantly more effort to build full systems. However, project partners are extremely experienced in this area and have carried out numerous related activities in the past. Therefore, we are confident that DIO will be able to deliver with the foreseen resources.

These resources and partner expertise cover all layers of the proposed architecture (Figure 5) both at the design as well as at the prototyping level. Other direct costs include travel, equipment, and audit certificates as discussed next. Finally, all partners have the internal structure and resources to support the project administratively.

Table 3.4a summarizes project effort across WPs and partners.

Table 3.4a: Summary of staff effort

	WP1	WP2	WP3	WP4	WP5	WP6	WP7	**PMs/ Partner**
1. FORTH	**12**	**54**	21	12	3	9	7	**118**
2. BSC	2	13	12	**48**	24	12	4	**115**
3. STFC	2	13	4	12	**24**	20	4	**79**
4. BULL	4	0	21	0	0	28	**10**	**63**
5. JGU	2	12	**30**	6	9	7	4	**70**
6. ICCS	2	0	6	6	6	**43**	4	**67**
7. CYB	2	0	7	0	18	8	6	**41**
Total PMs	**26**	**92**	**101**	**84**	**84**	**127**	**39**	**553**

3.4.3 Travel

Travel expenses are calculated for about 18 or 20 trips per partners for the duration of the project at about €1000 per trip. This is about 6 trips per year (per partner), 3 of which will be to project meetings and the additional 3 either for more people to participate in project meetings if required, for additional bi-lateral partner meetings for technical issues, or for dissemination and exploitation purposes. The coordinator (FORTH) has been allocated 24 trips and its travel budget includes the amount of €5000 for expenses of the advisory board. Finally, CYB is allocated 15 trips, as they are expected to have smaller needs due to the fewer full-time equivalents participating in the project.

3.4.4 Equipment

Carrying out the work proposed in this project will require extensive design, implementation, evaluation, and optimization. As mentioned above, partners have all the required expertise and capacity to carry out these tasks. Besides this expertise and staff resources, the project will use extensive infrastructure for implementation and evaluation purposes. This infrastructure involves state-of-the-art servers, storage devices and controllers and FPGA accelerators. However, all partners have already been involved in related activities, either research or product-oriented and thus, have to a large extent the required equipment in place. The needs for new equipment in this project will be small, and mostly extending existing systems either with newer or specialized components for specific cases and targeted measurements, as follows:

- **FORTH** already has two racks of servers for development and evaluation purposes of DIO with a lot of memory (256G/server) and fast interconnect. FORTH will require additional fast storage devices, accelerators, and possibly a few development servers for the purposes of the project, at a cost of €45000.
- **ICCS** will require about €20000 for FPGA accelerators for development and evaluation purposes.
- **BULL** will require about €30000 for upgrading its existing HPC solution dedicated to support R&D developments in DIO with updating computing nodes with Intel's skylake/cascadelake processor that can host NVM DIMMs, adding new acceleration technologies (FPGA), and fast storage capacity.

3.4.5 Other goods and services

Audit certificates

The estimated audit certificates required by project partners are: FORTH €3000, BSC €2625, JGU €4000, ICCS €3000, BULL €5000, STFC €5000 and the total amount reserved is €22625 (0.6% of project budget).

Other costs

FORTH includes in its other goods and services budget, in addition to the audit certificates, €3000 for publication expenses and €2000 for designing and printing project stationary.

Capitalised and operating costs of large research infrastructures

STFC will provide 340000 CPU hours on its large research infrastructure for the purposes of DIO in accordance with Article 6.2 of the General Model Agreement, as indicated in the table below.

3. STFC	Cost (€)	Justification
Large research infrastructure	25000	Cost for 340000 CPU hours on STFC infrastructure for development, testing, and evaluation purposes

Since no partner exceeds 15% of personnel costs for other direct costs, we omit Table 3.4b.

List of Selected Acronyms and Abbreviations

Term	**Description**	**Term**	**Description**
ANN	Artificial Neural networks	NVMe	Non-Volatile Memory Express
CFD	Computational Fluid Dynamics	PB	Petabytes
EB	Exabytes	PFS	Parallel File System
FPGA	Field-programmable gate array	PoC	Proof of Concept
GB	Gigabytes	SCM	Storage Class Memories
GPU	Graphics processing unit	SSD	Solid State Drive
HDD	Hard Disk Drive	SMTBF	System Mean Time Between Failures
HPDA	High Performance Data Analytics	TB	Terabytes
KV store	Key-value Store	TRL	Technology Readiness Levels
NVM	Non-Volatile Memory	UEABS	Unified European Applications Benchmark Suite

4 MEMBERS OF THE CONSORTIUM

4.1 Participants

4.1.1 FORTH, Greece

The Foundation for Research and Technology - Hellas (FORTH) is the premier research centre in Greece, internationally acknowledged for its excellence in basic and applied research, in developing applications and products, and in providing services. The Institute of Computer Science (ICS), one of the six Institutes of FORTH, will be contributing to this project. FORTH-ICS has a 30-year history of internationally competitive R&D contributions across the fields of Information & Communication Technologies, as well as of academic and industrial cooperation; it has adopted an evolving strategy to promote the commercial exploitation of R&D results by providing services, licensing products to industrial partners, contracting with industrial partners to jointly develop new products, and participating in spin- off companies and joint ventures. FORTH-ICS represents Greece in the European Research Consortium for Informatics and Mathematics (ERCIM). The Institute ranked first among all Greek institutes of its field in all national evaluation rounds so far (last evaluation 2014).

Role in the project

FORTH will co-ordinate the consortium, lead WP2 ("Data storage and access mechanisms"), and will work generally on issues related to key-value stores, memory-mapped I/O, RDMA-bsed data transfers, and checkpointing. The main tasks that FORTH will contribute with most of its effort and based on its expertise are T2.1 to T2.4, and T3.2.

Relevant Expertise

Contributions to this project will come from the Computer Architecture and VLSI Systems (CARV) Laboratory of FORTH-ICS that has a pool of more than 50 technical people; CARV has a 30-year history, in architecture, hardware, and systems software R&D, with fundamental contributions in high-speed communication, scalable computing, and, more recently, in server architectures, storage systems, programming models, and analytics infrastructure. CARV expertise includes the design, implementation, and test of dozens of innovative, real-life prototypes in both systems software and hardware.

In *Storage and I/O systems*, FORTH-ICS has worked on SSD caching, storage controller efficiency, scalability of the I/O path in modern servers, performance isolation in scalable multicore servers, distributed tier-0 storage for fast storage devices, and efficient higher-level abstractions for storage access. The work in CARV is experimental in nature and we have built numerous prototypes ,from storage controllers to distributed storage systems. Recently, and in the context of performance isolation, we have built an alternative, working I/O stack for the Linux kernel that achieves workload isolation for scalable servers. We are currently working on a rack-scale prototype for efficient access to storage for fast devices and analytics applications. Part of our work in CARV has been successfully commercialized between 2011-2013, through a large international company. In addition, our work during 2013-2016 has been licensed to a startup company, ioFabric, that has its main R&D in the campus of FORTH. Furthermore, FORTH is actively collaborating with Industry on real systems and products due to its expertise and track record on addressing problem in real systems.

In parallel architectures and interconnection networks FORTH-ICS worked, since 1986, on fair queuing, back-pressure, congestion tolerance, weighted round-robin scheduling and fair queuing, multiprocessor interconnects, and more; in 1996-98, a 6-million-transistor switch chip was designed and built. In Scalable Architectures, in the Telegraphos project (1993-95), workstation clustering prototypes were designed and built, including processor network interfaces for protected user-level communication, and in the last 8 years, on hardware and runtime systems for low-latency and high-bandwidth communication in chip multiprocessors, and in scalable parallel processing; a 520-core prototype has been designed and built, using 64 FPGA boards interconnected in a 3D mesh topology. Moreover, this line of work has resulted in FORTH having a main role in collaboration with Industry and academia in Europe in the design of ARM-based micro-servers. In this context, KALEAO Ltd, a European high-density server company has established its R&D lab in the campus of FORTH and actively collaborates with CARV. FORTH is an associate member of the ETP4HPC ("European Technology Platform in the area of High-Performance Computing") industry-led think-tank, where among other activities it has been active in the working group focusing on balancing compute, memory, and I/O requirements in next-generation HPC systems.

CARV has coordinated and participated in several EU projects. Currently, CARV is coordinating ExaNest (FETHPC) and participates with significant contributions in systems software, system protoyping and full-system evaluation in the

FETHPC projects ExaNoDe, Vineyard and EuroEXA, as well as in a number of various individual and collaborative national projects. Recently completed relevant projects (in the FP7 framework) include ASAP (where FORTH was the coordinator), LeanBigData, NanoStreams, and Euroserver. FORTH also participates in the ongoing BigStorage ETN, leading efforts towards the use of emerging NVM devices in the I/O stack.

Relevant Projects and Activities

CoherentPaaS [2013-2016]: Coherent and Rich PaaS with a Common Programming Model
During CoherentPaaS FORTH has designed Tucana, a key-value store that dramatically reduces overheads with respect to the state-of-the-art in the Cloud.
LeanBigData [2014-2017]: Ultra-Scalable and Ultra-Efficient Integrated and Visual Big Data Analytics
During LeanBigData FORTH has worked on efficient storage systems for datacenters aiming at low-overhead analytics. FORTH’s IP in LeanBigData has been licensed to industry. The knowledge acquired from solving the related problems and prototyping will be used to reduce overheads and embrace heterogeneity in the envisioned platform.
Vineyard [2016-2019]: Versatile Integrated Accelerator-Based Heterogeneous Data centers
Vineyard examines the integration of accelerators in datacentre servers in a manner that is transparent to applications and services. FORTH designs the transport between applications and accelerators that achieves accelerator sharing and virtualization. This work will be extremely valuable in incorporating acceleration in the envisioned platform.
EuroEXA [2017-ongoing]: A Hardware and Software Stack for Real-Time Analytics on Fast Data Streams
EuroEXA is a recently initiated H2020 project that is building an integrated and operational prototype with a rich mix of key applications from across the climate/weather, physics/energy and life-science/bioinformatics domains. FORTH is leading the systems software work-package, and is the main contributor and maintainer of the operating system and firmware for the three successive generations of the project's rack-scale testbeds.
ExaNeSt [2015-ongoing]: European Exascale System Interconnect and Storage
ExaNeSt, a FETHPC project coordinated by FORTH, is developing rack-scale prototypes that support memory and I/O resource sharing (according to the UNIMEM architectural paradigm), and aims to port, tune and evaluate a diverse set of real HPC and data-management Applications. FORTH is a key contributor in the work-packages focused on system interconnects and storage, actively supporting the porting and optimization efforts of application owners. In the area of storage in particular, FORTH has been developing a low-latency memory-mapped storage access path in the Linux kernel, as well as a fast path for protected accesses to a key-value store.

Relevant Publications, Products, and Services

1. Anastasios Papagiannis, Giorgos Saloustros, Manolis Marazakis, and Angelos Bilas. 2017. Iris: An optimized I/O stack for low - latency storage devices. SIGOPS Oper. Syst. Rev. 50, 3 (January 2017), 3-11.
2. Anastasios Papagiannis, Giorgos Saloustros, Pilar Gonzalez-Ferez, and Angelos Bilas. Tucana: Design and Imlementation of a Fast and Efficient Scale-up Key-value Store. In proceedings of the 2016 USENIX Annual Technical Conference (USENIX ATC'16). June 2016, Denver, CO, USA.
3. Nikolaos Papakonstantinou, Foivos S Zakkak, Polyvios Pratikakis. Hierarchical Parallel Dynamic Dependence Analysis for Recursively Task-Parallel Programs. International Parallel and Distributed Processing Symposium (IPDPS), Chicago, IL, USA, 2016.
4. Pilar Gonzalez-Ferez and Angelos Bilas. Reducing CPU and network overhead for small I/O requests in network storage protocls over raw Ethernet. In Proceedings of the 31st International Conference on Massive Storage Systems and Technology (MSST’2015), Santa Clara, CA, USA, June 2015.
5. Yannis Sfakianakis, Stelios Mavridis, Anastasios Papagiannis, Spyridon Papageorgiou, Markos Fountoulakis, Manolis Marazakis, and Angelos Bilas. Vanguard: Increasing server efficiency via workload isolation in the storage i/o path. In Proceedings of the 2014 ACM Symposium on Cloud Computing (SoCC’14), Seattle, WA, USA, November 2014.
6. Vineet Bafna, Alin Deutsch, Andrew Heiberg, Christos Kozanitis, Lucila Ohno-Machado, and George Varghese. Abstractions for genomics. Commun. ACM56, 1 (January 2013), 83-93.

7. Thanos Makatos, Yannis Klonatos, Manolis Marazakis, MichailD. Flouris, and Angelos Bilas. Using Transparent Compression to Improve SSD-based I/O Caches. In Proc. of The EuroSys 2010 Conference (Eurosys'2010), April 2010.

Infrastructure

The CARV lab within FORTH has available for work in this direction an enterprise datacenter-level rack with:

- CPU: #cores = 512
- Main Memory: amount of memory (DRAM): 2.5 TBytes
- Interconnect: 40GBit/s Ethernet+Infiniband
- Storage: amount of storage: 4TB SSD, 10TB disk

During the project this equipment will be enhanced with additional components (server boards, controllers, accelerators, storage) for specific experiments related to the work in the project.

Key Personnel for the Project

Angelos Bilas (male) is currently a Professor of Computer Science at FORTH-ICS and the University of Crete, Greece, where he has also held an Associate Professor position between 2002-2011. Prof. Bilas received his diploma in Computer Engineering from the University of Patras in 1993, and the M.S and Ph.D. degrees in Computer Science from Princeton University, NJ in 1995 and 1998 respectively. Between 1998-2002 he held an Assistant Professor position with the ECE Department at the University of Toronto. His current interests include architectures and efficient systems software for datacenter servers, server architectures, storage systems, low-latency high-bandwidth communication protocols, and runtime-system support for multicore processors. His work has been published in prestigious conferences in computer architecture and systems. Prof. Bilas is the recipient of a Marie Curie Excellent Teams Award (2005-2009), has served in the Program Committees of more than 80 conferences and workshops in his area, as the Program Chair for IEEE Cluster 2010, in the Editorial Board of IEEE Computer Architecture Letters (CAL) [2007-2011], currently serves in the Editorial Board of the Journal of Parallel and Distributed Computing (JPDC) [May 2011-], has participated in more than 15 EU or nationally funded projects, and was the coordinator of the FP7 EU project IOLanes [2010-2013]. [http://www.ics.forth.gr/~bilas]

Dr. Manolis Marazakis (male) is a Staff Research Scientist at FORTH-ICS. He received his Ph.D. in 2000 in Computer Science from University of Crete. He works on architectures and efficient systems software for high-performance servers and storage systems. He has contributed to the design, implementation and performance evaluation of several system prototypes, including online transparent compression for storage systems, SSD caching, storage controller efficiency, efficiency and scalability of the I/O path in modern multicore servers, storage area networks, and low-latency I/O prcoessing for storage hosts and virtual machines. See: http://www.ics.forth.gr/~maraz

Dr. Christos Kozanitis (male) is a postdoctoral scholar at FORTH-ICS. He received his M.S. and Ph.D in Computer Science and Engineering from the University of California, San Diego in 2009 and 2013 respectively. He held a two-year postdoctoral appointment at the AMP Lab of the University of California, Berkeley, where he used and adapted state of the art big data technologies, such as Apache Spark SQL, Apache Parquet and Apache Avro to process large amount of DNA sequencing data. His current research interests involve the improvement in software, storage and hardware level of modern datacenters in order to speed up the processing of big data workloads.

4.1.2 BSC, Spain

The Barcelona Supercomputing Center (BSC) was established in 2005 and is the Spanish national supercomputing facility and a hosting member of the PRACE distributed supercomputing infrastructure. The Center houses MareNostrum, one of the most powerful supercomputers in Europe. The mission of BSC is to research, develop and manage information technologies in order to facilitate scientific progress. BSC was a pioneer in combining HPC service provision, and R&D into both computer and computational science (life, earth and engineering sciences) under one roof. The centre fosters multidisciplinary scientific collaboration and innovation and currently has over 400 staff from 41 countries. In 2011, BSC was one of only eight Spanish research centres recognized by the national government as a “Severo Ochoa Centre of Excellence”.

BSC has collaborated with industry since its creation, and has participated in projects with companies such as ARM, Bull and Airbus as well as numerous SMEs. BSC also participates in various bilateral joint research centers with companies such as IBM, Microsoft, Intel, NVIDIA and Spanish oil company Repsol. The centre has been extremely active in the EC

Framework Programmes and has participated in over one hundred projects funded by it. BSC is a founding member of HiPEAC, the ETP4HPC and participates in the most relevant international roadmapping and discussion forums and has strong links to Latin America.

Education and Training is a priority for the centre and many of BSCs researchers are also university lecturers. BSC offers courses as a PRACE Advanced Training Centre, and through the Spanish national supercomputing network among others. Computer Sciences: The BSC-CNS Computer Sciences Department focuses on building upon currently available hardware and software technologies and adapting these technologies to make efficient use of supercomputing infrastructures. The department proposes novel architectures for processors and memory hierarchy and develops programming models and innovative implementation approaches for these models as well as tools for performance analysis and prediction.

Role in the project

BSC will lead WP4. The main role of BSC in DIO will be the scheduling, sharing, and isolation aspects (WP4) and NEMO application (WP5). All these areas tie extremely well with the expertise available at BSC and will also support in the best possible manner future research directions in the storage group within BSC**.**

Relevant Expertise

Personnel of BSC-CNS taking part to the project are members of the Computer Sciences Department of the BSC-CNS. More precisely, they are members of the Storage System group that has an extensive experience with data management in all levels. The group works from the node level to wide-area distributed level. In the node level, the group has contributed on new file systems and optimizations of the I/O stack. At the cluster level, several contributions on cooperative caches as well as heterogeneous and scalable storage systems have been achieved. Finally, the StorageSystem Group has worked on replica management and location in XtreemFS, which is a wide-areafile system proposed in the context of the EU XtreemOS IP Project. Finally, BSC has worked on replica management and location in XtreemFS, which is a wide-area file system. All this work has been awarded by a number of European and National projects (Paros, Nanos, POP, XtreemOS FP6, HPC1-5, VELOX FP7, IOLanes FP7, NEXTGenIO H2020, IOStack H2020).

Relevant Projects and Activities

NEXTGenIO [2015-2018]: Next Generation I/O Project
BSC is leading the system ware work page and building several file systems (object and file based) to use efficiently the NVM.
ESiWACE [2015-2019]: Centre of Excellence in Simulation of Weather and Climate in Europe.
The goal is to substantially improve efficiency and productivity of numerical weather and climate simulation on high-performance computing platforms by supporting the end-to-end workflow of global Earth system modelling in HPC environment.
PRIMAVERA [2015-2019]: PRocess-based climate sIMulation: AdVances in high resolution modelling and European climate Risk Assessment
The main objective is to develop a new generation of advanced and well-evaluated high-resolution global climate models, capable of simulating and predicting regional climate with unprecedented fidelity, for the benefit of governments, business and society in general.
IOLanes [2010-2013]: Advancing the Scalability and Performance of I/O subsystems in multi-core Platforms
During IOLanes BSC created methods to analyse the workload and select the best I/O scheduler automatically. Additionally, data deduplication techniques with low overhead in the kernel had been developed.
MontBlanc 2 [2013-2017]: European scalable and power efficient HPC platform based on low-power embedded technology.
BSC researched a way to reduce writes in RAID 5 and RAID 6 systems, leading to a performance improvement.

Relevant Publications, Products, and Services

1. A. Miranda, S. Effert, Y. Kang, E.L. Miller, A. Brinkmann, T. Cortes. Reliable and Randomized Data Distribution Strategies for Large Scale Storage Systems. 18th Annual International Conference on High Performance Computing Bangalore, India, December 18-21, 2011.
2. Paul Hermann Lensing, Toni Cortes, André Brinkmann. Direct Lookup and Hash-Based Metadata Placement for Local File Systems. 6th Internationa System and Storage Conference (Systor 2013) Haifa, Israel, Jun 30-July 2, 2013.
3. R. Nou, J Giralt, T.Cortes. DYON: Managing a New Scheduling Class to Improve System Performance in Multicore Sys-tems. 1st Workshop on Runtime and Operating Systems for the Many-core Era (ROME 2013) Aachen, Germany, August 26, 2013.
4. R. Nou, J Giralt, and T. Cortes. Automatic I/O scheduler selection through online workload analysis. 9th IEEE International Conference on Autonomic and Trusted Computing Fukuoka, Japan, September 4-7, 2012
5. Ramon Nou, Alberto Miranda and Toni Cortes. Performance impacts with Reliable Parallel File Systems at Exascale level. Europar 2015, Vienna, Austria, August 24-28, 2015.

Infrastructure

BSC hosts and will provide access for testing purposes to the following infrastructures:

- MareNostrum 4 will have a performance capacity of 13, 7 Petaflop/s. This innovative supercomputer will be made from IBM, which will integrate in one unique machine its own technologies alongside those of Lenovo, Intel and Fujitsu. The general purpose element, provided by Lenovo, will have 48 racks with 3,456 nodes with next generation Intel Xeon processors and a central memory of 390 Terabytes. Its peak power will be over 11,1 Petaflop/s, which is to say that it will be able to perform more than 11,000 trillion operations per second while the extra power will be achieved using heterogeneous and new hardware like Intel KNH. It will have an Elastic Storage of 15 PBytes.
- MinoTauro - a cluster with 128 Bull B505 blades, with 1,536 processors, 256 M2090 NVIDIA GPU cards, 3,07 GB of main memory.

Key Personnel for the Project

Prof. Toni Cortes (male) is the manager of the storage-system group at the BSC (since 2006) and is also an associate professor at Universitat Politècnica de Catalunya (since 1998). He received his Ph.D. in computer science in 1997 (at Universitat Politècnica de Catalunya). Since 1992, Toni has been teaching operating system and computer architecture courses at the Barcelona school of informatics (UPC) and from 2000 to 2004 he also served as Vice Dean for international affairs at the same school. His research concentrates in storage systems, programming models for scalable distributed systems, and operating systems. He has published 26 journal papers, 76 papers in international conferences and workshops. In addition, he has also advised 10 PhD theses since 1997. Dr. Cortes has been involved in several EU projects (Paros, Nanos, POP, XtreemOS, SCALUS, IOlanes, PRACE, MontBlanc, IOStack, BigStorage and NextGenIO) and has also participated in cooperation with IBM (TJW research lab) on scalability issues both for MPI and UPC.

Ramon Nou (male) has been working at BSC (since 2006) on the Autonomic System and e-Business Platforms group of BSC until 2008 when he switched to the Storage-System group as a researcher. He has been working on SORMA, XtreemOS, IOLanes, MontBlanc, PRACE-2IP, BigStorage, IOStack and NEXTGenIO EU Projects. In 2008, he obtained his Ph.D. with Prof. Jordi Torres as advisor with the topic "Using online simulations to improve QoS on middleware". He has published more than 20 papers in international conferences and workshops, and has three journal papers. Ramon has a wide view on all computer levels, with expertise on optimization, performance measurements and simulation/modelling of complex systems. Ramon has also knowledge on Bioinformatics and Biostatistics and is pursuing a master's degree on the topic.

Alberto Miranda (male) has been working since 2007 as a researcher in advanced HPC storage systems for the storage-system group at the Barcelona Supercomputing Center. Dr. Miranda received a diploma in Computer Engineering, a M.S. degree in Computer Science and a M.S. degree in Computer Architectures, Networks and Systems from the Technical University of Catalonia in 2004, 2006 and 2008, respectively. He received a Ph.D. degree Cum Laude in Computer Science from the Technical University of Catalonia in 2014. His current research interests include efficient file and storage systems, operating systems, distributed systems architectures, as well as information retrieval systems. He is also deeply interested in the applications of machine learning and automated learning systems in all these areas. He has published 4 papers in international conferences and 1 journal paper. Dr. Miranda has been involved in EU-funded projects XtreemOS, IOLanes, Prace2IP, IOStack, Mont-Blanc 2, EUDAT2020, Mont-Blanc 3, and NEXTGenIO.

4.1.3 STFC Hartree Centre, United Kingdom

The Science and Technology Facilities Council (STFC) is one of the UK's seven publicly fundedResearch Councils responsible for supporting, co-ordinating and promoting research, innovation and skills development in seven distinct fields. STFC is a world-leading multi-disciplinary science organisation combining access to large-scale facilities with funding for research and innovation.

- Universities: STFC supports university-based research, innovation and skills development in astronomy, particle physics, nuclear physics, and space science.
- Scientific Facilities: STFC provides access to world-leading, large-scale facilities across a range of physical and life sciences, enabling research, innovation and skills training in these areas.
- National Campuses: STFC works with partners to build National Science and Innovation Campuses based around our National Laboratories to promote academic and industrial collaboration andtranslation of our research to market through direct interaction with industry.
- Inspiring and Involving: STFC helps ensure a future pipeline of skilled and enthusiastic young peopleby using the excitement of our sciences to encourage wider take-up of STEM subjects in school and future life (science, technology, engineering and mathematics).

STFC supports an academic community of around 1,700 in particle physics, nuclear physics, and astronomy including space science, who work at more than 50 universities and research institutes in the UK, Europe, Japan and the United States. STFC's main facilities are located at two UK campuses: the Rutherford Appleton Laboratory at Harwell in Oxfordshire and the Daresbury Laboratory in Cheshire. Currently, STFC employs around 1700 members of staff in addition to funding over 900 PhD students.

Following significant capital investments from UK government, the Hartree Centre was established to offer a range of services including collaborative software development and access to a range of novel hardware platforms. The Hartree Centre combines strengths in business development with skills and expertise from the Scientific Computing Department to create world-class multi-disciplinary capabilities, reinforced by deep ties with industry and academia. The technical expertise in hardware and software development for HPC and scientific expertise in an extensive range of applications codes, together with the provision of a range of HPC resources, will be crucial to this project.

Role in the project

Tha Hartree Centre will lead WP5 ("Application adaption and evolution") and will also significantly contribute based on its experience towards delivery of WP6 ("Acceleration, integration, and evaluation of the converged scalable architecture), especially extrapolation of the obtained results in this project on the future exascale architectures and WP4 ("Resoruce and QoS management"). In general tha Hartree Centre's main role will be in co-simulating hardware and software for future exascale architectures, non-instrusive in-depth analysis and adaption of applications and near-data computing by utilitising technologies such as RDMA. Tasks from this work packages fit extremly well with the expertise available in the Hartree Centre and the future direction of the centre.

Relevant Expertise

The work undertaken in this project will be by the members of The Future Technologies (FT) group within the Hartree Centre and Scientific Computing Department (SCD).

The members of FT group have a comprehensive and deep expertise in hardware/software co-design through implementation of high-level full-system architectural simulator as well as full understanding of data movement between multiple levels of storage hierarchies. The group has several contributions towards quantifying in-depth memory traffic between storage devices (such as HDD and SDD), caches and processors and use of this feedback to schedule parallel tasks more efficiently performance- and energy-wise such that time and energy use is minimized during data movement.

The SCD Department has substantial experience in developing, maintaining, debugging, porting and benchmarking several codes on different platforms, i.e. IBM Blue Gene/L/P/Q, IBM POWER clusters, Cray XT4, XT6, XE6, XC30, Bull clusters. In the last 10 years, they further developed their HPC experience, mainly due to involvement in PRACE projects and lately performed a simulation over 3 million of process cores on Mira, Argonne's IBM Blue Gene/Q using

Code_Saturne, an open-source CFD software, within the INCITE PEAC project. They are now intensively working with the EDF Energy UK R&D centre simulate the flow and temperature distribution in EDF Energy Advanced-Gas cooled Reactor fuel assemblies. They have performed the largest production simulation to date, using Code_Saturne, with a 1.6 billion cell mesh, within an ESPRC ARCHER Leadership Award.

Relevant Projects and Activities

Project List
COMPAT [2015-2018]: Computing Patterns for High Performance Multiscale Computing
COMPAT is a sceince driven project whose aim is to create a general mapping from multiscale model to computing patterns that can effectively run on future exascale machines. The part of STFC in this project is to provide middleware services that are capable of efficient execution of those patterns on various resources.
PRACE 4iP [2015-2017]: The Fourth Implementation Phase project
STFC was involved in updating Code_Saturne's test cases UEABS in order to have the benchmark ready for a set of data to be tested on Tier 0, 1, and 2 systems. Furthermore, extension of this project allowed measurement of energy consumption on PCP machines.
DCIM [2015-2018]: Data Centre Infrastructure Management Project
This project implemented a state of the art data centre management system that provides the capability to measure power, energy, temperature, and humidity measurements across all machine rooms at various levels of granularity. These power and energy consumption readings are correlated with the application execution data stored in the central database to support infrastructure management and optimisation as well as validating more granular energy measurements
Vineyard [2016-2019]: Versatile Integrated Accelerator-Based Heterogeneous Data centers
Vineyard examines the integration of accelerators in datacentre servers in a manner that is transparent to applications and services. In this project STFC provides a scheduling framework that is used to find the best available HPC systems that will run the application in the most efficient manner both performance- and energy-wise by minimizing amount of memory traffic that needs to be transferred.
EuroEXA [2017-2020]: Co-designed Innovation and System for Resilient Exascale Computing in Europe: From Applications to Silicon
EuroEXA is a recently initiated H2020 project that is building an integrated and operational prototype with a rich mix of key applications from across the climate/weather, physics/energy and life-science/bioinformatics domains. The role of STFC in this project is to provide necessary infrastructure in order to provide test bed where extrapolation of work towards exascale machines will be performed.

Relevant Publications, Products, and Services

1. C. Moulinec, D.R. Emerson, Y. Fournier, P. Vezolle, "Challenges to be Overcome for Engineering Software to Run Efficiently on Petascale Machines", in B.H.V. Topping and P. Iványi, (Editor), "Developments in Parallel, Distributed, Grid and Cloud Computing for Engineering", Saxe-Coburg Publications, Stirlingshire, UK, Chapter 2, pp 23-40, 2013
2. DJ Greaves, M Puzovic, AM Zaidi, K McDonald-Maier, Andrew Hopkins. “Fine-grained Energy/Power Instrumentation for Software-level Efficiency Optimization” at the Forum on Description Languages, FDL'15, Barcelona, 14th -16th September 2015.
3. C. Moulinec, J. C. Uribe, J. Gotts, B. Xu, D. R. Emerson. “Sleeve leakage gas impact on fuel assembly temperature distribution”. International Journal of Computational Fluid Dynamics Vol. 30 , Iss. 6, 2016
4. M Puzovic, S Manne, S GalOn, M Ono: Quantifying Energy Use in Dense Shared Memory HPC Node. E2SC@SC 2016: 16-23
5. David Topping, Irfan Alibay, and Michael Bane, “Accelerating activity coefficient calculations using multicore platforms, and profiling the energy use resulting from such calculations”, EGU2017-12246N. Di Pasquale, M. Bane, S.J. Davie and P.L.A. Popelier (2016), "FEREBUS: Highly Parallelized Engine for Kriging Training", J. Comput. Chem., vol. 37, 2606-2616.“Proceedings of the EMerging Technology (EMiT) Conference 2016”, Editors: B.D.Rogers, D.Topping, F. Mantovani, M.K.Bane. ISBN 978-0-9933426-3-9.
6. D.O. Topping, M. Barley, M. Bane, N.J. Higham, B. Aumont, and G. McFiggans (2016), “UManSysProp: An online facility for molecular property prediction and atmospheric aerosol calculations”, Geosci. Model. Dev. 9, pp899-914.
7. DJ Greaves, M Puzovic: Prazor/VHLS User Manual. https://goo.gl/svq6Fj

Infrastructure

The Hartree Centre (STFC) will provide access for testing purposes to the following infrastructure:

- Scafell Pike has a perfomrnace capacity of 3.4 PFLOPs. It consists of 846 dual Intel Xeon E5 2699v5 with 128HB memory, 840 nodes Intel Xeon Phi 7210 64 core with 96 GB memoyr, 24 High Memory Nodes with 2 Intel Xeon E5 2699v5 and 1TB memory, 30 data hierarchy nodes with Intel Xeon Phi 64 core processors, 384GB high performance memory and local NVMe drive and 16 data mover and 16 interactive nodes, where each node is configures with 2 Intel Xeon E52688v5 and 126GB high performance memory as well as additional network interface.

Key Personnel for the Project

Dr Milos Puzovic (male) is a Research Scientist at the Hartree Centre. His main field of research is the optimisation of software for performance and power consumption through hardware, operating system, compiler and run-time system co-design. In addition to work on software optimisation he is also working in the area of the full-system architectural simulation in order to study in-depth microarchitectural design space trade-offs. Milos has completed his PhD at the University of Cambridge on the subject of hardware and software co-design for dynamic multicore scheduling. He has supervised undergraduate courses in operating systems, optimising compilers and comparative architectures and has experience from working in finance and computer software industries.

Dr Michael Bane (male) is a Research Scientist at the Hartree Centre. Previously Bane managed the Research Applications Team within IT Services at The University of Manchester ensuring support of researchers in their computational and data analytic needs whether desktop or high end compute, ensuring an institution-wide RDM strategy, and personally developed the highly acclaimed research computing training. Bane has decades of experience in optimizing and parallelizing scientific codes. Bane is Chair of the Emerging Technology (EMiT) conference series. His current research at the Hartree Centre involves measuring, predicting and minimizing energy-to-solution of software applications on various hardware platforms.

Dr Charles Moulinec (male) works at Scientific Computing Department in Daresbury Laborttory since 2007 and is a Principal Scientist since 2013. He obtained his doctorate in CFD from Ecole Centrale de Nantes (France) in 1996 and then moved to the Netherlands (TU Delft) to work as a PDRA to improve gradient reconstruction for very skewed grids (1996-98) and then to perform DNS of tube bundles using Cartesian grids (1998-2002). From 2003 to 2005, he worked as a PDRA (KTP scheme between UMIST-The University of Manchester and CD adapco) on LES for grids made of polyhedral cells. He has been working in computational fluid dynamics (CFD) problems for more than 20 years.

4.1.4 Atos/BULL, France

Atos/BULL is a large IT company that specializes in the design and development of servers and HPC systems from hardware up to application level. The Group, which is firmly established in the Cloud and in Big Data, integrates and manages high-performance systems (including high end HPC Clusters) and end-to-end security solutions. Bull's offerings enable its customers to process all the data at their disposal, creating new types of demand. Bull converts data into value for organizations in a completely secure manner. Bull currently employs around 9,200 people across more than 50 countries, with over 700 staff totally focused on R&D. In 2016, Bull recorded revenues of €855 M€ with a particularly strong presence in the public, healthcare, finance, telecommunications, manufacturing and defense sectors.

Closely attuned to changes in technology and business, Bull R&D is involved in all key areas of the Group’s activities: high-performance computing, cyber-security, Cloud computing, information systems modernization, Big Data. In recent years, the Bull R&D labs have developed many major products that are recognized for their originality and quality. These include the bullx supercomputer offer (with several references in the TOP500), bullion servers for the private Clouds and Big Data, the Shadow intelligent jamming system designed to counter RCIEDs, the libertp tool for modernization of legacy applications and, most recently, hoox, the first European smartphone featuring native security. To explore new areas and develop tomorrow's solutions, today, Bull R&D is investing heavily in customers – with whom it has forged many successful technological partnerships – as well as in institutional collaborative programs (such as competitiveness clusters and European projects) and in partnerships with industry (Open Source, consortiums).

Role in the project

Bull will lead the Dissemination and Exploitation activities in the project (WP7), will participate with significant technical effort to the design and validation of a Hierarchical Checkpointing Protocol for heterogeneous computing systems in WP3, in integrating accelerators in the IO path (WP6), and will participate in system evaluation (WP6).

Relevant Expertise

Bull is particularly active in the application sectors of defense, finance, health care, manufacturing, public and telecommunication for which it offers a wide range of in-depth skills in both hardware design (system architecture, ASIC, electronic boards, communications, ect.) and software development (middleware, system/application doftware).

Relevant Projects and Activities

Project List (5 max)
Catrene/TSAR
Design, development and prototyping of a many-core (1024) processor
Catrene/SHARP
Design, and Prototyping of a hypbid HPC solution to mix a variety of computing technologies including GPGPU, FPGA and many-core processor TSAR.
Vineyard [2016-2019]: Versatile Integrated Accelerator-Based Heterogeneous Data centers
Vineyard examines the integration of accelerators in datacentre servers in a manner that is transparent to applications and services. FORTH designs the transport between applications and accelerators that achieves accelerator sharing and virtualization. This work will be extremely valuable in incorporating acceleration in the envisioned platform.
Network-based HPC BXI
A large internal project at Bull dedicated to the development of a network-based HPC solution including the design of many ASICs (NIC, router) and the development of middleware/software (Portals, MPI, etc.).
Node Controller xNC
Design of a node-controller ASIC (+ BIOS and firmware development) dedicated to the building of next generation of 32-sockets servers at Bull

Relevant Publications, Products, and Services

1. Ania Kaci, Huy-Nam Nguyen, Amir Nakib, Patrick Siarry, "Hybrid Heuristics for Mapping Task Problem on Large Scale Heterogeneous Platforms", IEEE Computer Society , vol. 00, no. , pp. 809-816, 2016, doi:10.1109/IPDPSW.2016.163
2. Huy-Nam Nguyen, Tuan-Anh Nguyen "Transaction-Level Simulation of Network-based Computing Systems" IP SoC 2013
3. Huy-Nam Nguyen “Specification & Validation of Cache Coherence protocols” DTC 2010
4. Huy-Nam Nguyen, Eric Guthmuller “Prototyping of a scalable, cache coherent, shared memory, multi-cores architecture” Tutorial “ESL Design and Virtual Prototyping of MPSOCs” at DAC 2010

Infrastructure

BULL has available data centers providing access to HPC systems with hardware acceleration technologies.

Key Personnel for the Project

Huy-Nam Nguyen (male) is an engineer in automation. He received his Docteur-Ingénieur in Applied Mathematics. Since 1981 he is with Bull where he participates to the design and development of many generation of proprietary or open servers. His topics of interest include, ASIC design and (Formal) Verification, Hardware Emulation and System Rapid Prototyping. In the last decade, he represents Bull in many European co-operation projects including Jessi/AC8, Medea/AT403-AT407, Medea+/A502-A511-2A718 Catrene/TSAR, Catrene/SHARP and has contributed actively to the Medea+ Eda roadmap (Eds 2003, 2005 and 2008).

Patrice Bulot (male) is an engineer in computer sciences. He participates to many generation of proprietary and open servers at Bull as a specialist in hardware emulation and system prototyping. He also contributed to the French cooperation project FUI/WASGA Server and the European Catrene/SHARP project.

4.1.5 Johannes Gutenberg University Mainz (JGU), Germany

JGU is one of the largest universities in Germany and hosts more than 32,500 students from about 120 nations. As the only comprehensive university in Rhineland-Palatinate, JGU combines almost all academic disciplines, including the Mainz University Medical Center, the School of Music, and the Mainz Academy of Arts. With 75 fields of study and a total of 242 degree courses, including 106 Bachelor's and 116 Master's degree programs, JGU offers an extraordinarily broad range of courses. Some 4,360 academics, including 560 professors, teach and conduct research in JGU's more than 150 departments, institutes, and clinics. JGU is a globally renowned research university of national and international recognition.

This reputation comes thanks to its outstanding individual researchers as well as extraordinary research achievements in the field of particle and hadron physics, materials sciences, translational medicine, the life sciences, media disciplines, and historical cultural studies. JGU is one of 23 universities in Germany that have received approval for a so-called Cluster of Excellence as well as approval for a Graduate School of Excellence. The Cluster of Excellence on "Precision Physics, Fundamental Interactions and Structure of Matter" (PRISMA), which is primarily a collaboration between particle and hadron physicists, and the Graduate School of Excellence "MAterials Science IN MainZ" (MAINZ) are considered among the elite research groups worldwide. These two projects will receive financing to the tune of EUR 50 million by 2017.

Role in the project

JGU will lead WP3, and its main focus will be the DIO APIs. It will port legacy APIs (T3.1) and design novel, storage-optimized APIs (T3.3 in conjuction with T2.2). It will also provide support to all workloads of WP4 and WP5 to use the APIs of WP3.

Relevant Expertise

JGU participates in the project through the "Zentrum für Datenverarbeitung" (ZDV). ZDV is the university's data centre and the focal point for HPC and storage related research activities. The personnel at ZDV consist of more than 60 technicians and 12 Ph.D. students and post-docs. ZDV is hosting a number of high-performance computers, including Mogon II with 16,500 Cores and a performance of 557 TFlops. JGU is represented by the ZDV as a full member in the German Gauß Alliance. ZDV focuses its research on storage systems, Cloud computing, and HPC. Previous storage systems research includes scalable storage virtualization, storage architectures, file systems, energy efficient storage systems, data deduplication, and archiving. Cloud computing expertise includes resource management systems, energy efficient scheduling, service level agreements, and malleable applications.

Storage system related research at the ZDV is always performed in the context of scalable and efficient systems. The development cycle is based on the concept of algorithm engineering, where each cycle typically starts with an investigation of the underlying physical and algorithmic problems in scale-out systems, continues with the design and the analysis of parallel and distributed algorithms and also includes the overall system design and its integration.

The investigation of existing challenges is typically performed together with international partners, e.g., the ECWMF in UK, the Lawrence Livermore National Labs in the USA, the Barcelona Supercomputing Center in Spain, or the DKRZ in Germany. The collaboration with these partners allows us to investigate huge parallel environments, disclosing real world challenges on the way to Exascale computing, where gathered traces are made available as open data if possible (see, e.g., https://www.ecmwf.int/en/en/computing/our-facilities/data-handling-system).

Application contexts include the design of flash translation layers (and their implementation) on the device level to provide application-specific storage solutions, the design of multi-process OS-bypass libraries to support an efficient access to storage class memory, kernel extensions for storage virtualization, the development of scalable data deduplication environments, the design and implementation of parallel file systems and quality of service mechanisms. Several of the implementations have been made available through industry collaborations, now being, e.g., part of the standard Lustre release.

Relevant Projects and Activities

[redacted]
DFG ADA-FS [2016-2019]: Advanced Data Placement via Ad-hoc File Systems at Extreme Scales
ADA-FS aims to improve I/O performance for highly-parallel applications by distributed ad-hoc overlay file systems. For this purpose, it examines how job-specific temporary file systems can be efficiently provided for HPC environments. These file systems are to be created from the resources of the computing nodes involved. The temporary file systems are

filled with the necessary data through an integration into the scheduling system of the supercomputer before the job starts. After the completion of the job, the data is migrated back into the global parallel file system. The research approach includes both the design of the file system itself as well as the questions about the proper scheduling strategy for planning the necessary I/O transfers. The project is related to this proposal, as virtualized environments have to build up on demand and with low resource overhead in both projects.

DFG SFB / TRR 146 [2014-2018]: Multiscale Simulation Methods for Soft Matter Systems

The ZDV coordinates the central support project within the SFB / TRR 146, which offers software development services for the science projects and which is itself a research project that analyses and optimizes the usage of HPC resources within the context of soft-matter systems. The coordination project provides, e.g., an asynchronous checkpointing environment, which is able to scale by distributing the checkpoints in the HPC cluster, using techniques, which are independent from the programming framework. It also allows applications to steer their IO-behavior and therefore enables them to improve IO performance even after the development process has finished. Especially the checkpointing techniques can be partially transferred to DIO.

Intel® Parallel Computing Center for Lustre [2016-2017]: Lustre QoS: Network Request Scheduler and Monitoring Revisited

The Network Request Scheduler (NRS) has been introduced in Lustre's mainline kernel in version 2.4.0 to provide different request scheduling options. The NRS has been extended (partly in this project) to offer QoS techniques including an enhanced token bucket strategy, where an average bandwidth is assigned to each client. The "Lustre QoS"-project will include additional information to improve the quality of the NRS and to optimize overall bandwidth delivery. The main idea is to include information about the striping targets of each client into the token bucket strategy to ensure that no individual OST will be overloaded. General QoS techniques can also be transferred to the DIO project.

EU BigStorage [2013-2017]: Converegence of HPC and Big Data and Storage Tiering

BigStorage is a European Training Network (ETN) to train future data scientists to enable them and us to apply holistic and interdisciplinary approaches for taking advantage of a data-overwhelmed world, which requires HPC and Cloud infrastructures with a redefinition of storage architectures underpinning them - focusing on meeting highly ambitious performance and energy usage objectives. The PhD students at JGU focus on new storage technologies like storage class memory (SCM), on energy-efficient storage, and on developing unified frameworks to support HPC and Big Data applications.

Relevant Publications, Products, and Services

5. Fabio Margaglia, Gala Yadgar, Eitan Yaakobi, Yue Li, Assaf Schuster, André Brinkmann: The Devil Is in the Details: Implementing Flash Page Reuse with WOM Codes. In Proceedings of the 14th USENIX Conference on File and Storage Technologies (USENIX FAST), Santa Clara, CA, USA, pages 95 – 109, February 22-25, 2016
6. Giuseppe Congiu, Sai Narasimhamurthy, Tim Süß, André Brinkmann: Improving Collective I/O Performance Using Non-volatile Memory Devices. In Proceedings of the IEEE International Conference on Cluster Computing (IEEE CLUSTER), Taipei, Taiwan, pp. 120 – 129, September 12-16, 2016
7. Paul Hermann Lensing, Toni Cortes, Jim Hughes, André Brinkmann: File System Scalability with Highly Decentralized Metadata on Independent Storage Devices. In Proceedings of the IEEE/ACM 16th International Symposium on Cluster, Cloud and Grid Computing (IEEE/ACM CCGrid), Cartagena, Colombia, pp. 366 – 375, May 16-19, 2016
8. Jürgen Kaiser, Tim Süß, Lars Nagel, André Brinkmann: Sorted deduplication: How to process thousands of backup streams. In Proceedings of the 32nd Symposium on Mass Storage Systems and Technologies (MSST), Santa Clara, CA, USA, pp. 1 – 14, May 2-6, 2016
9. Fabio Margaglia, André Brinkmann: Improving MLC flash performance and endurance with extended P/E cycles. In Proceedings of the IEEE 31st Symposium on Mass Storage Systems and Technologies (IEEE MSST), Santa Clara, CA, USA, pp. 1 – 12, May 30 - June 5, 2015

Infrastructure

MOGON I and MOGON II: The ZDV runs the HPC systems MOGON I and II and provides access to them for all researchers in Rhineland-Palatinate and their collaborators. MOGON I has been purchased in 2012 and reached number 81 in the Top 500 ranking of the fastest academic supercomputers. It includes 555 nodes and provides 35,520 AMD bulldozer cores, running at a Linpack performance of more than 250 TFlop/s. The 825 nodes of the new HPC system

MOGON II have been deployed end of 2016 and the 16,500 Xeon E5-2630v4 10C 2.2GHz cores achieve a Linpack performance of 557 TFlop/s.

Key Personnel for the Project

Prof. Dr.-Ing. André Brinkmann is a full professor at the computer science department of JGU and head of the ZDV (since 2011). He received his Ph.D. in electrical engineering in 2004 from the Paderborn University and has been an assistant professor in the computer science department of the University of Paderborn from 2008 to 2011. Furthermore, he has been the managing director of the Paderborn Centre for Parallel Computing PC^2 during this time frame. His research interests focus on the application of algorithm engineering techniques in the area of data centre management, cloud computing, and storage systems. He has published more than 100 papers in renowned conferences and journals and is an associated editor of the ACM Transactions on Storage as well as a steering committee member of the IEEE Symposium on Massive Storage Systems and Technologies (MSST) and IEEE International Conference on Networking, Architecture, and Storage (NAS). Prof. Brinkmann is a member of the advisory board of the French Grid'5000 and of the EU Horizon 2020 Modular Microserver DataCentre (M2DC) project.

Dr. Tim Süß got his PhD from the Paderborn University and is currently a post-doctoral research assistant at the Johannes Gutenberg University Mainz. He received his diploma (M.Sc.) degree in computer science in 2007 from the Paderborn University. His current research interests are energy aware computing, storage systems, and scheduling in the context of HPC. Beside his research activities he is also involved in the administration of the university's HPC cluster MOGON and the guidance of its users.

4.1.6 Institute of Communications and Computer Systems, Greece

The Institute of Communications and Computer Systems (ICCS) is a non-profit Academic Research Body established in 1989 by the Ministry of Education in order to carry research and development activities in the fields of all diverse aspects of telecommunications and computer systems. ICCS is associated with the School of Electrical and Computer Engineering (SECE) of the National Technical University of Athens (NTUA). The personnel of ICCS consists of a number of Research scientists and more than 500 Associate scientists (including PhD students). The research carried out in ICCS is substantially supported by School of Electrical and Computer Eng. University Professors. ICCS is very active in European co-funded research activities and has been the project manager of many EU projects in various programs in all of the above mentioned research areas. The Microprocessors and Digital Systems Laboratory (MicroLab) of ICCS, which will participate in the current proposal, is actively involved in the implementation of digital systems, from system level up to chip level, using the most advanced and modern technologies and methodologies, ranging from microprocessors and microcontrollers, embedded systems development boards, signal processors, special purpose ICs and FPGAs, to the actual design of VLSI Application Specific ICs (ASICs). At the moment, μLab consists of 23 scientists (2 professors, 6 research associates and 15 Ph.D. students). The MicroLab has been granted many European and National Research Programmes (> 55 projects) regarding with the embedded systems and VLSI systems design.

Role in the project

ICCS will lead WP6, and will work generally on issues related to the hardware acceleration of data processing and data movement. Also ICCS will lead the efforts for the implementation of prototype platfroms based on FPGA technology that will allow the storage devices (e.g. NVMe) to be connted to the network through the FPGA devices in order to increase the throughput and reduce the latency of the starge systems.

Relevant Expertise

ICCS has a long track record, more than 15 years, in developing design- and run-time techniques for the optimization of embedded systems based on digital systems design of multicore/many-core as well as the design and implementation of accelerators of computationally-intensive applications onto reconfigurable systems realised in data centers, space rovers & debris and biomedical equipment. More specifically, the Microlab has experience on various aspects of the reconfigurable & embedded systems design including middleware services, design space exploration, memory management, resource management in terms of performance, power consumption, memory footprint, dependability and reliability criteria. The main results of all projects have been published in relevant international conferences and journals, and, received best paper awards and the EU-sponsored HiPEAC awards. Additionally, ICCS has long experience on the coordination and execution of plethora dissemination activities in several FP7 and H2020 projects.

Regarding with memory management field, MicroLab has activities more than ten years mainly in the field of Embedded Computing and High Perfomance (Embedded) Computing. Design Methodologies, Simulators, design techniques, several publications in journals, conferences and books, European (FP5-AMDREL, FP7-MNEMEE, FP7-MOSART, FP7-

2PARMA) and National funded projects cover main aspects of memory management field. The delivered memory management material took/takes into account many design criteria, such as, perfomance, power consumption, energy mangement and dependability. In particular, research activities and innovative results on design space exploration, efficient data layout and physical memory organization for customized/optimized memory hierarchies under different computing platforms (e.g., microprocessors, FPGAs, DSP processors) performed in R&D projects. Additionally, a couple of patents from members of MicroLab are granted.

Relevant Projects and Activities

Vineyard [2016-2019] Versatile Integrated Accelerator-Based Heterogeneous Data centers

ICCS is the project coordinator of the H2020 VINEYARD project. VINEYARD's goal is to develop the technology and the ecosystem that will enable the efficient integration of the hardware acceleration in the data centre applications, seamlessly. The deployment of energy-efficient hardware accelerators will be used to improve significantly the performance of cloud computing applications and reduce the energy consumption in data centres. VINEYARD is developing an integrated framework for energy-efficient data centres based on programmable hardware accelerators. It is working towards a high-level programming framework that allows end-users to seamlessly utilize these accelerators in heterogeneous computing systems by using typical data-centre cluster frameworks (i.e. Spark). The VINEYARD framework and the required system software hides the programming complexity of the heterogeneous computing system based on hardware accelerators.

VINEYARD also foster the establishment of an ecosystem that will empower open innovation based on hardware accelerators as data-centre plugins, thereby facilitating innovative enterprises (large industries, SMEs, and creative start-ups) to develop novel solutions using VINEYARDS's leading edge developments. The ecosystem will bring together existing communities from all relevant stakeholders including providers of hardware intellectual-property (IP) technologies, data centre developers, data centre operators and more. This ecosystem will allow the promotion of open pluggable custom hardware accelerators (i.e. a hardware accelerator Application Store) that can be used in data centres in the same way that software libraries are currently being utilized.

TOISE [2011-2013]: Trusted Computing for European Embedded devices, No. 270001-2, Funding Sceme ENIAC-2010-1, JTI-CP-ENIAC Project Coordinator Bernard Candaele, Thales, France

The objective of TOISE is to define, develop and validate trust hardware and firmware mechanisms applicable both to lightweight embedded devices and as security anchors within related embedded platforms. ICCS developed a systematic methodology for designing custom memory manager to deal with multi-core trusted and embedded platforms. More specifically, the customized dynamic memory allocators have a specific layer/component that offers an adaptive fit policy, as the application(s) needs change at run-time. The data management techniques take into considerations the constraints imposed by the adaptive task allocation and power management techniques.

MNEMEE [2008-2010]: *M*emory ma*n*ag*eme*nt technology for adaptive and efficient design of *e*mbedded systems," 7th IST Framework, No. 216224, STREP, Objective: *ICT-2007.3.3: Embedded Systems Design*

The goal of the MNEMEE project was to develop source-to-source optimization methodologies and tools to improve the design of MPSoC embedded systems, realizing ambitious future applications. The developed optimizations make possible the mapping of very demanding applications and increase the cost efficiency of the final mapping. The optimization methodologies and tools are divided in two parts: i) compiler-independent part and ii) memory-hierarchy aware (but processor architecture independent) part. ICCS Provided a framework for source-to-source optimization methodologies, which targets both statically and dynamically allocated data of complex embedded software applications. Statically allocated data means that all the data is stored in memory in the beginning of the execution time (e.g. arrays). Dynamically allocated data means that the data is stored and removed from memory during run-time (e.g. single-linked lists), in order to adapt to changing conditions and user needs. Also, ICCS was the leader of WP2.

MOSART [2008-2010]: Mapping Optimisation for Scalable multi-core ARchiTecture," 7th IST Framework, STREP, Objective: *ICT-2007.3.4 Computing systems a) Novel architectures for multi-core computing systems*.

The goal of the MOSART project was to develop a flexible, modular, multi-core on-chip platform architecture and associated exploration design methods and tools. The developed approach allows the scaling of the platform and optimisation of its constituent elements for various embedded, multimedia and wireless communication applications.

ICCS developed data management methods, middleware and runtime support for optimization of memory accesses, data transfers, allocation and assignments. Also, ICCS was the leader of WP2.
2PARMA [2010-2013]: PARallel PAradigms and Run-time MAnagement techniques for Many-core Architectures, STREP FP7-ICT-2009.3.6-4, (1/1/2010 - 31/3/2013). Additional comments: (i) Ranked as 1st proposal of 40 submitted projects and (ii) EU considered 2PARMA as "Success Story".
The 2PARMA project aimed at overcoming the lack of parallel programming models and run-time resource management techniques to exploit the features of many-core processor architectures. ICCS was actively involved in the resource management techniques and implementations by designing a dynamic memory allocator / manager. Extensive effort has been put on taking the proper decisions during both the design- and run-time of the manager. The manager has been validated on a number of many-core platforms, including a general-purpose (x86) server platform and an embedded many-core accelerator (STHORM). Finally, ICCS organized the dissemination's activity for the academia side of the project.

Relevant Publications, Products, and Services

1. E. Koromilas, I. Stamelos, C. Kachris, D. Soudris, **Spark acceleration on FPGAs: A use case on machine learning in Pynq**, 2017 6th International Conference on Modern Circuits and Systems Technologies (MOCAST), pp. 1-4, 2017
2. C. Kachris, D. Soudris, **A survey on reconfigurable accelerators for cloud computing**, 2016 26th International Conference on Field Programmable Logic and Applications (FPL), FPL 2016
3. C. Kachris, G. Sirakoulis, D. Soudris, **A Reconfigurable MapReduce Accelerator for multi-core all-programmable SoCs**, International Symposium on System-On-a-Chip, (SOC'14), Tampere, Finland, October 2014
4. C. Kachris, G. Sirakoulis, D. Soudris, **A Configurable MapReduce Accelerator for Multi-core FPGAs**, 22nd ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA'14), Monterey, CA, Feb. 2014
5. C. Kachris, G. Sirakoulis, D. Soudris, **Performance Evaluation of Embedded Processor in MapReduce Cloud Computing Applications**, International Conference on Cloud Computing (CloudComp'12), Wien, Austria, Sept. 2012
6. D. Atienza, J.M. Mendias S. Mamagkakis, D. Soudris, and F. Catthoor, "**Systematic Dynamic Memory Management Design Methodology for Reduced Memory Footprint**," in *ACM Transactions on Design Automation of Electronic Systems* (TODAES) Vol. 11, No. 2, April 2006, Pages 465–489

Key Personnel for the Project

Dr. Dimitrios Soudris (male) received his Diploma in Electrical Engineering from the University of Patras, Greece, in 1987. He received the Ph.D. Degree in Electrical Engineering, from the University of Patras in 1992. He was working as Lecturer, Ass. and Assoc. Professor in Dept. of Electrical and Computer Eng., Democritus University of Thrace for thirteen years since 1995. He is currently working as Assoc. Professor in School of ECE of National Technical University of Athens, Greece. His research interests include embedded systems design, reconfigurable architectures, network-on-chip architectures and low-power VLSI design. He has published more than 300 papers in international journals and conferences. Also, he is author and editor in eight books of Kluwer and Springer. His research work has been cited >1000 times. He is the Head of Embedded Systems group consisting of four Senior Investigators (Post-Docs) and eight PhD students and a number of M.Sc. students. He is leader and principal investigator in numerous research projects (>45) funded from the European Commission, ENIAC-JU, European Space Agency and the Greek Government and Industry. He has served as General Chair and Program Chair for PATMOS and General Chair of IFIP-VLSI-SOC 2008, PARMA 2011 & 2013 Workshop, Chair of Track "Modeling, Design, and Design Space Exploration Track" of SAMOS XIV Conference 2014. Also, he served and serves Technical Program Committees of many International Conferences. He received an award from INTEL and IBM for the project results of LPGD #25256 and awards from Int. Conf. VLSI 2005 and ASP-DAC 05 for the results of the project AMDREL IST-2001-34379 and recently the HiPEAC Award for a DAC '10 paper for the project MNEMEE FP7-216224 and DAC'13 for the project ENIAC-TOISE-282557-2. He will contribute his long experience in research and development of embedded systems and be the overall technical manager of ICCS.

Dr. Christoforos Kachris (male) received his Ph.D. in computer engineering in 2007 from the Technical University of Delft, The Netherlands in 2008 and the diploma and the M.Sc. in Electronic and Computer Engineering from the Technical University of Crete, Greece in 2001 and 2003 respectively. From 2009 to 2010 he was a visiting lecturer at University of Crete and a visiting researcher at FORTH where he coordinated the Interconnects cluster of High

Performance Embedded Architectures and Compilers (HiPEAC). In 2006 he was a research intern at Xilinx Research Labs, San Jose, CA, working at the Networks Group. He has participated in many European Projects in the domain of high performance embedded systems for telecommunications such as the SARC IP project, CHRON, NAVOLCHI, ACCORDANCE, and the COCONUT project. He is member of the HiPEAC NoE on embedded systems since 2007. He has published more than 50 papers in international journals and conferences and he was the editor of the book "Optical Interconnects for Future Data Center Networks". His main research interest is in the area of embedded systems, reconfigurable computing (FPGAs), high-speed network processing, multicore embedded systems, IoT, computer architecture, and interconnects.

4.1.1 CYBELETECH SAS

CybeleTech is a young SME, established in 2011, that aims at developing the use of numerical technologies in agriculture. The core products of CybeleTech are based on either numerical simulations of plant growth through dedicated biophysical models or machine learning methods extracting knowledge on processes through large databases.

These new technologies can bring added values at different stages of the agriculture and food chain:

- In plant breeding, where simulating plant growth can help reducing the amount of field trials by approximately 50% and consequently reduce the time needed to produce a new cultivar.
- For optimization of cultural practices, helping farmers to save resources and to maximize yields.
- Forecasting yields and production at large scale in order to better anticipate storage and market variations.
- For the optimization of first transformation processes, helping to guarantee the best output quality from agricultural products with heteregoneous quality.

CybeleTech has tied links with academic partners, including INRA (French National Institute for Agronomic Research) for soil and plant modeling and databases, Ecole Centrale Paris for plant growth modeling and CEA for sensors development and usage.

Role in the project

CybeleTech will work on the evaluation of the prototype with an HPDA application working on large scale earth observation data for agricultural production forecast. The main implications of CybeleTech will be in WP 5 for adapting the workflow of the application and in WP 6 for evaluation of the prototype and extrapolation to large scale systems. CybeleTech will also help in WP 3 to the definition of I/O library support for image processing applications.

Relevant Expertise

CybeleTech has user-side experience on the development of large-scale parallel applications in support of its research and development activities. The company has tied links with the HPC community for the development and support of its applications. For instance, CybeleTech benefited of the support of PRACE through a SHAPE project, a program dedicated to SMEs, in order to optimize the workflow and scalability of its parallel applications. CybeleTech is a member of ETP4HPC since 2015.

CybeleTech has strong expertise on applications working on large datasets of earth observation imaging, in particular processing data from the Sentinel network. Such applications are developed based on machine learning methods with high potential for parallelization and HPDA in various areas including deep learning and data assimilation in models.

Relevant Projects or Activities

- Smart Agriculture System: National project for optimization of nitrogen fertilization on wheat with earth observation data. In collaboration with Limagrain, Telespazio, John Deere, AgroPithiviers, Ecole Centrale Paris and Chambre d'Agriculture du Loiret. The project will end in december 2017.
- PALM (Product of Agriculture Lifecycle Management): National project for nitrogen strategies and yield potential on corn, rapeseed and wheat. In collaboration with Limagrain, Axereal, Ecole Centrale Paris, Université d'Orléans. The project will end in december 2018.
- SHAPE project (PRACE) on plant breeding: 400.000 core hours were awarded to compute numerical simulations of plant growth in order to optimize measurement protocols for seed companies.

- MAGESTAN: National project for automated management of greenhouses, working on the demonstrator case of tomatoes. In collaboration with CTIFL, INRA, Wi6labs. The project will end in december 2018.

Relevant Publications, Products, and Services

1. 2015: Prize for start up of the year at the "Trophy of numerical simulations" by l'Usine Nouvelle and Ter@tec.
2. B. Cirou, **D. Fernandez**, G. Hautreux, **D. Wouters**; Parallelization and optimization for plant selection with CybeleTech (2016) PRACE Highlights
3. AgreenTech Valley in Orléans: creation of an association gathering industrial and academic partners of the agri-food chain for incubation of innovative projects. Building of a physical campus in Orléans.
4. "Plant breeding goes numeric", article in "Semences et Progrès" N°177 (January 2017)

Key Personnel for the Project

Diane Fernandez (female), is project manager at CybeleTech with strong expertise in statistical methods for data analysis and assimilation as well as image processing. She has worked on the modeling of nitrogen processes in a soil-plant-atmosphere system and the assimilation of earth observation data in order to reduce model uncertainties during the season. At CybeleTech, she is the leader of the national project PALM, in collaboration with Limagrain and Axereal, working on the optimization of nitrogen fertilization for farmers. She holds a PhD in fundamental physics from University of Montpellier.

Denis Wouters (male), is the research and development lead at CybeleTech. He has strong expertise in the domain of machine learning technologies and high performance computing. He holds a PhD in fundamental physics from University Paris Sud 11 (Orsay).

4.2 Third parties involved in the project (including use of third party resources)

Partner BSC provides the following tabular information at the request of the European Commission. For all other partners the statement that applies is: “no third parties will be used”.

Partner BSC	
Does the participant plan to subcontract certain tasks (please note that core tasks of the project should not be sub-contracted)	N
If yes, please describe and justify the tasks to be subcontracted	
Does the participant envisage that part of its work is performed by linked third parties	N
If yes, please describe the third party, the link of the participant to the third party, and describe and justify the foreseen tasks to be performed by the third party	
Does the participant envisage the use of contributions in kind provided by third parties (Articles 11 and 12 of the General Model Grant Agreement)	Y
Some of the work carried out at the Barcelona Supercomputing Center – Centro Nacional de Supercomputación will be contributed free of charge by Third Parties: Universitat Politècnica de Catalunya (UPC), the Catalan Institution for Research and Advanced Studies (ICREA), and the Spanish Council for Scientific Research (CSIC). The BSC is a consortium that is composed of the following member institutions: Universitat Politècnica de Catalunya (UPC), Spanish Council for Scientific Research (CSIC), as well as the Spanish and the Catalan governments. Both UPC and CSIC contribute in kind by making human resources available to work on projects. The relationship between BSC and CSIC / UPC (respectively) is defined in an agreement with each institution that was established prior to the start of this project. **Universitat Politècnica de Catalunya (UPC)** The High Performance Computing research group of the Computer Architecture Department at the Universitat Politècnica de Catalunya (UPC) is the leading research group in Europe in topics related to high performance processor architectures, runtime support for parallel programming models, performance tuning applications for supercomputing and Cloud Computing. Directly derived from the research effort at the Computer Architecture Department, the CEPBA (European Center for Parallelism in Barcelona) was founded in 1991 to offer supercomputing resources to the research community and as a development center for industrial computing technology products. In 2000, IBM joined forces with CEPBA to form the CIRI (CEPBA-IBM Research Institute Joint Lab) in Barcelona in order to strengthen relationships between IBM and UPC researchers in computer architecture. In 2005, the Spanish and Catalan governments signed an agreement with IBM to buy the 4th supercomputer in the world and extend the operations of CIRI to become the Barcelona Supercomputing Center (BSC). The High Performance Computing research group at the UPC shares many key resources with the BSC, including several key personnel that will be dedicated to this project. There is a signed Collaboration Agreement between the UPC and the BSC establishing the framework of the relationship between these two entities. According to this agreement, several professors of the UPC are made available to the BSC to work on projects.	

5 ETHICS AND SECURITY

5.1 Ethics

This proposal is concerned with basic data storage technologies that will improve the cost-effectiveness and capabilities of our storage infrastructures. As such, there are no ethical issues.

5.2 Security

Co-ordinator FORTH has co-ordinated and participated in FP7 Security Theme projects where there are significant numbers of materials classified at RESTREINT-UE and CONFIDENTIEL-UE levels. Based on this experience FORTH provides the following answers on behalf of the DIO consortium.

Please indicate if your project will involve:

- Activities or results raising security issues: **NO**
- 'EU-classified information' as background or results: **NO**

[End of Document]

European Commission

PARTICIPANT
PORTAL
E-RECEIPT

Digitally signed by sealingservice.grants.ec.europa.eu
Date: 2017.09.26 18:24:51 CEST
Reason: Acknowledgment of Receipt