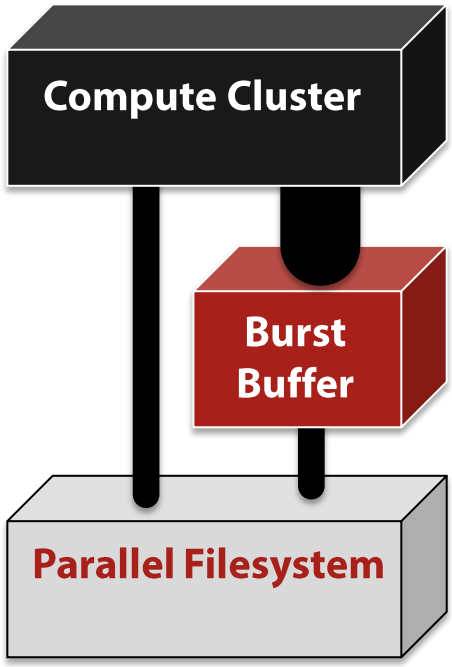
Application Performance on IME

Toine Beckers, DDN
Marco Grossi, ICHEC

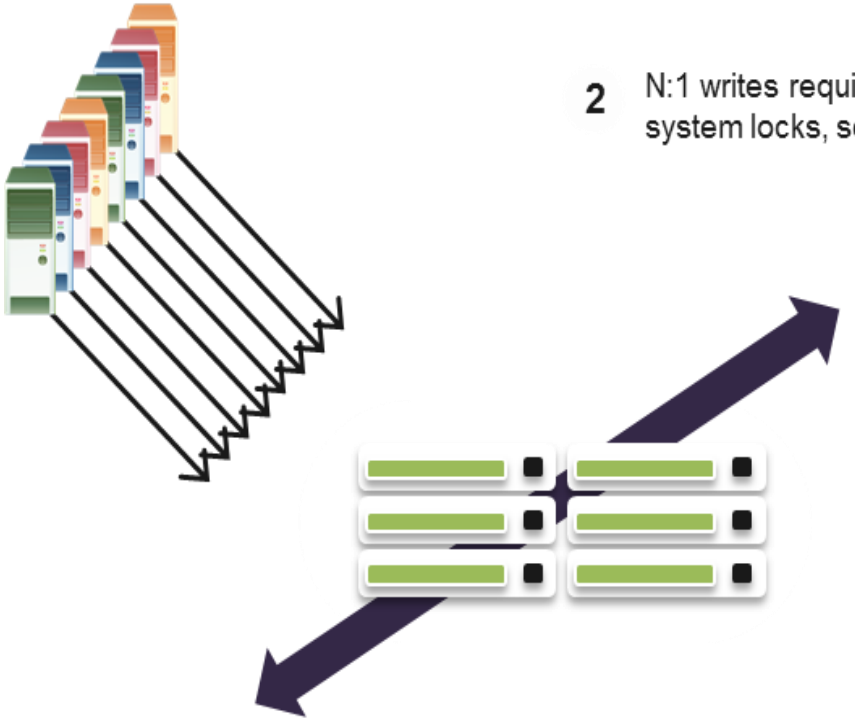
Burst Buffer Designs



- **Introduce fast buffer layer**
- **Layer between memory and persistent storage**
  - Pre-stage application data
  - Buffer writes from memory to fast devices
  - Store intermediate application data
- **Still a "mount point" (similar to a file system)**

3 Infinite Memory Engine: How does it Work?

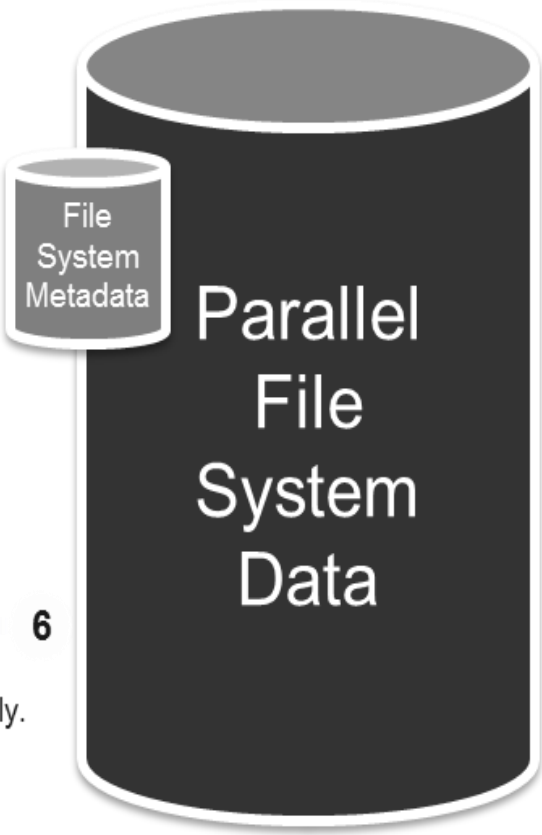
1 Exascale applications Issue a parallel 'burst' write request to IME



2 N:1 writes require no parallel file system locks, scale completely linearly

3 Writes are dynamically mapped to the buffer based upon load and node health

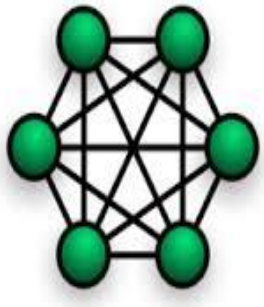
5 Data is selectively "drained" to a parallel file system in a completely sequential, non-contentious manner, almost near-perfect disk utilization



6 Persisted reads are staged either into the buffer or to the client directly. Because the write was 100% sequential, reads are equally fast

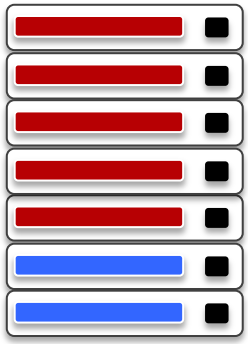
4 Data is protected across IME resources to withstand any node or media failure

IME Summary



Designed for Scalability
Ultra-low latency I/O between Compute Nodes and NVM

Fully POSIX & HPC Compatible
Additional APIs Available



Scale-Out Data Protection
Distributed Erasure Coding

Non-Deterministic System
Write Anywhere, No Layout Needed

Integrated With File Systems
Accelerates Lustre, GPFS
No Code Modification Needed

Writes Fast; Read Fast Too
No other system offers both at scale

ICHEC Background

- **Irish Centre for High-End Computing**
  - National Technology Centre
  - Established in 2005 → 10th anniversary!
- **Powered by people**
  - 27 staff
  - Terrific mix of computational scientists, researchers, developers and systems administrators
  - Dublin(east coast) & Galway(west coast) office
- **Mandates include**
  - HPC & Big Data/Data Analytics
  - Industry engagement
  - Partnerships, consultancy, training & services
  - Public sector & agency engagement
  - Services, enablement & training
  - National Academic HPC Service
  - Collaboration, training & service provision

TORTIA Intro

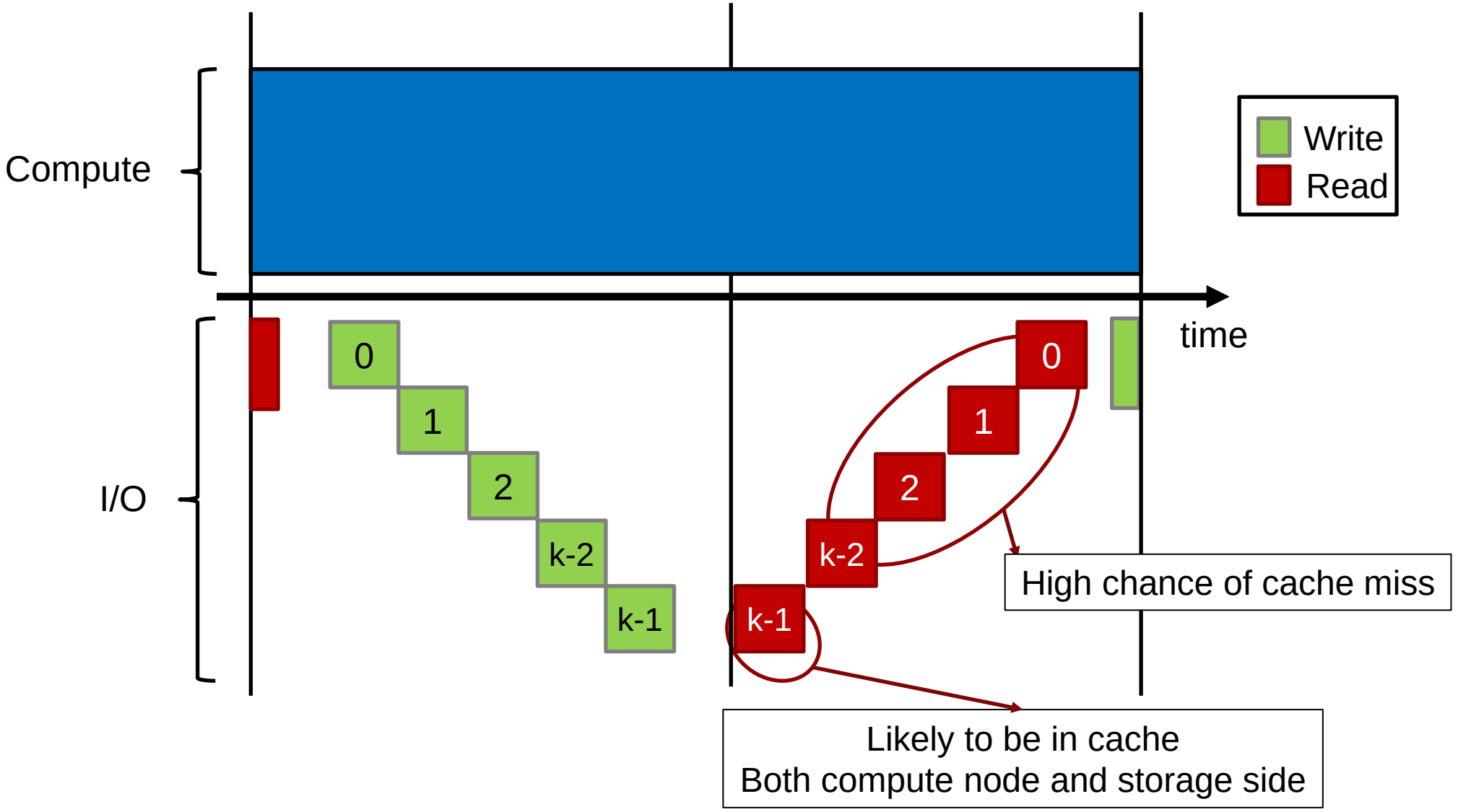
- **TORTIA (Tullow Oil Reverse Time Imaging Application)**
  - Developed in house for, and in collaboration with, Tullow Oil plc
- **A real application for real work!**
- **Reverse Time Migration (RTM) code**
  - Used by Oil & Gas companies to analyse seismic survey data
- **TORTIA is heavily optimized and tuned**
  - Parallelism, vectorization, … but also optimized on the I/O side
  - Achieves 30-50% of peak at scale

TORTIA
Some details

- **Standard C++ with OpenMP & MPI**
- **Input and output data in SEG-Y format**
- **Requires a temporary scratch area**
  - First half of the time loop dump snapshots of velocity fields
  - The second half of the time loop read back the saved snapshots
  - LIFO (Last-In First-out) access pattern
- **Implement 3 different I/O backend for the scratch**
  - POSIX
  - MPI-IO
  - In Memory aka “no I/O”

TORTIA
Scratch I/O pattern: LIFO

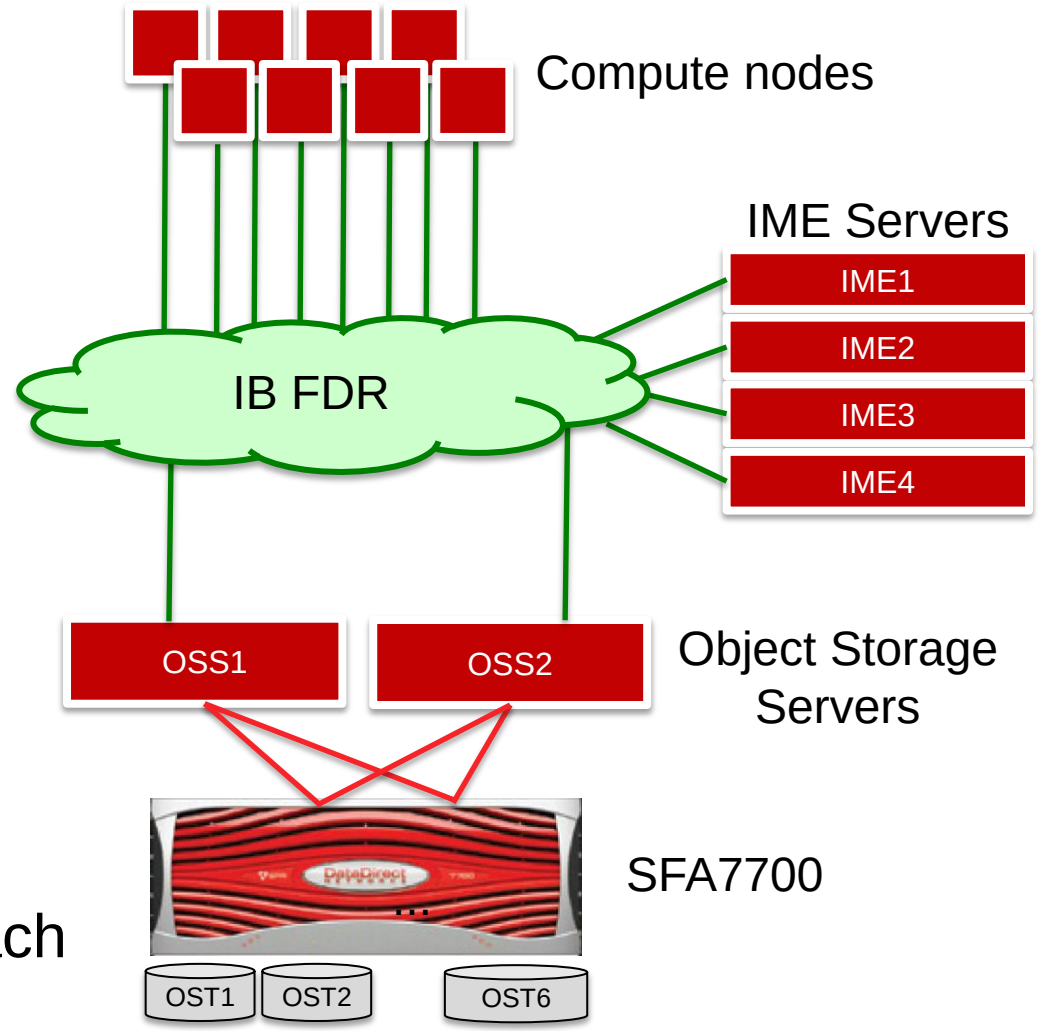
ICHEC
Irish Centre for High-End Computing



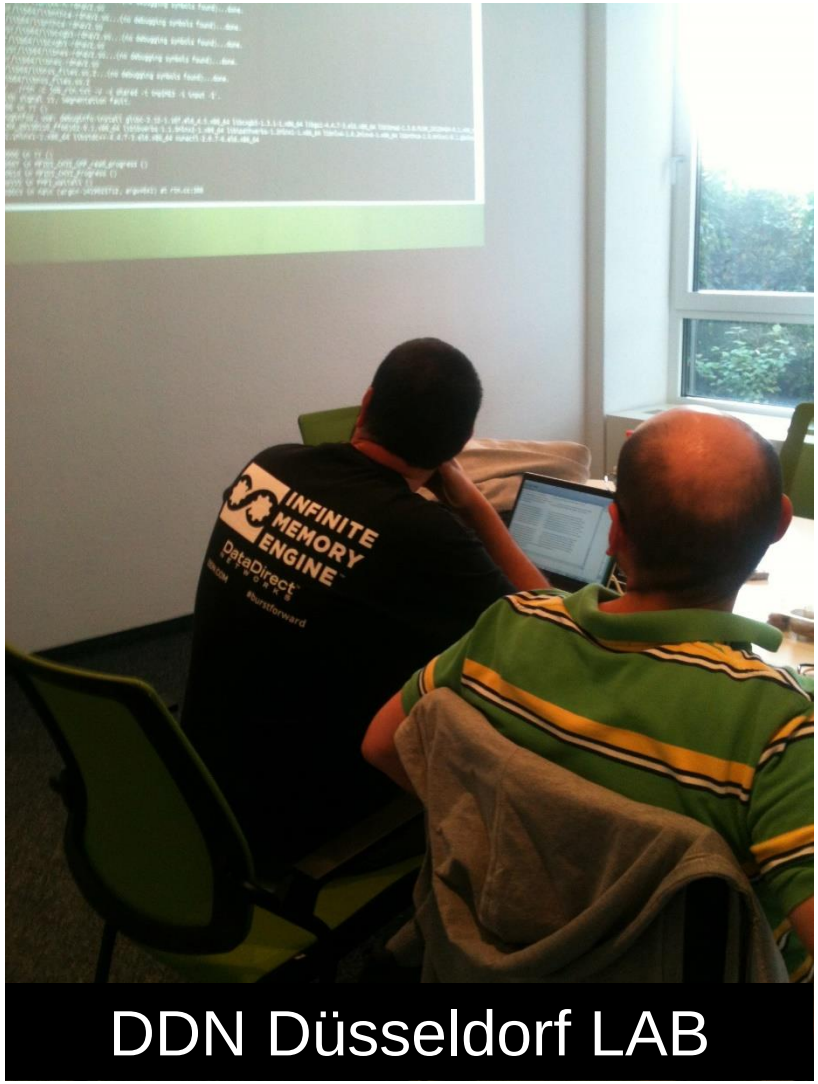
TORTIA on pre-GA DDN IME Test cluster

ICHEC
Irish Centre for High-End Computing

- **8 x Compute Nodes**
  - 2x Intel Xeon E5-2680v2
  - 128GB RAM
  - FDR InfiniBand
- **Filesystem Storage**
  - DDN SFA 7700
  - Lustre 2.5 with 2 x OSS servers
  - 3.4GB/s Write, 3.3 GB/s Read
- **IME System**
  - 4 servers with 24 x 240GB SSDs each
  - 36GB/s Write, 39 GB/s Read



TORTIA
Code porting



DDN Düsseldorf LAB

- **Used the MPI-IO interface to DDN IME**
- **Some constraints on IME pre-GA**
  - Required patched version of MVAPICH2
  - Added IME libraries at link time
- **Prepended ‘im:’ to file path**
- **Used MVAPICH instead of Intel MPI**
  - Still used Intel Compiler

TORTIA
Experiment use case

ICHEC
Irish Centre for High-End Computing

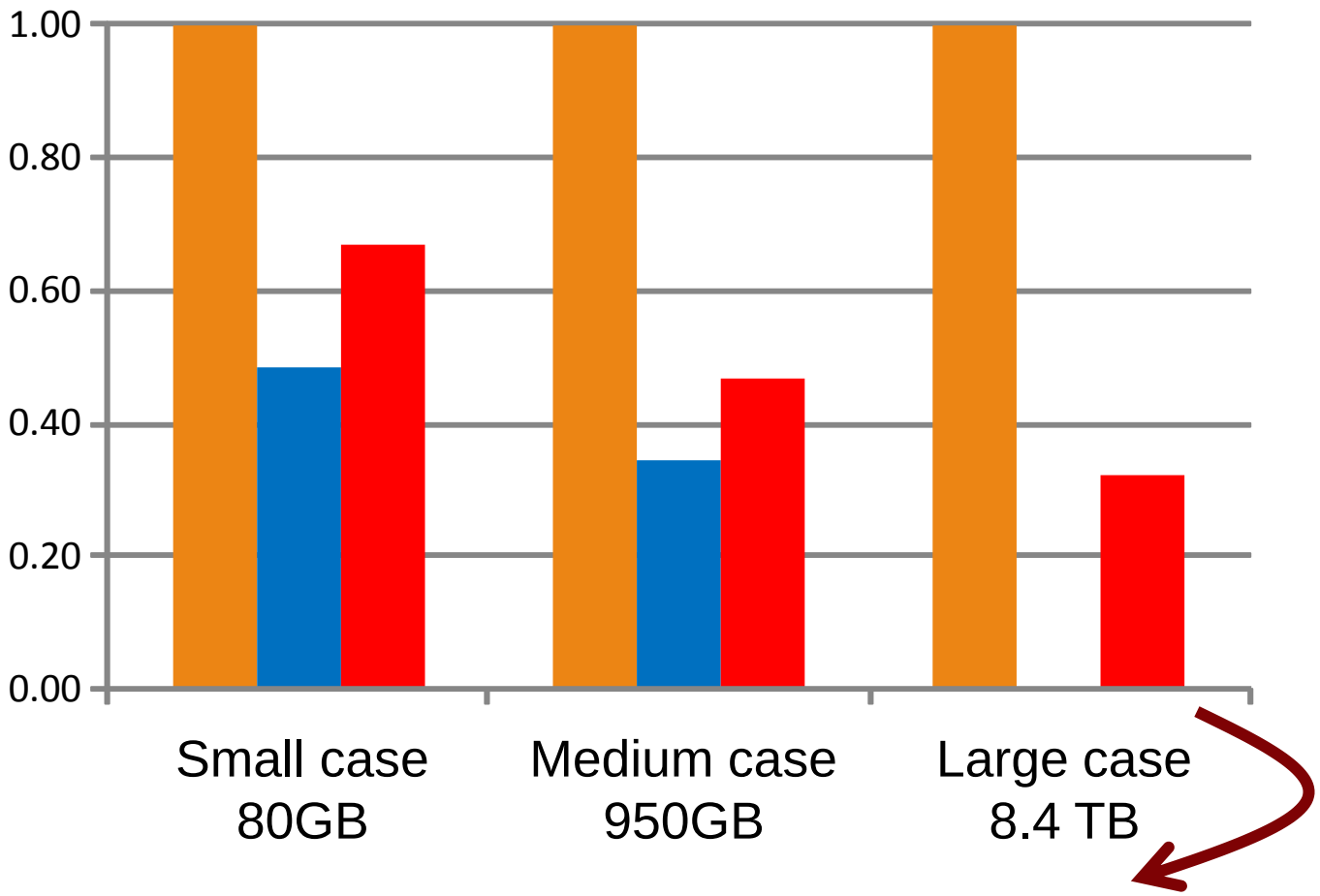
Scratch I/O target	Interface
In-memory	-
Lustre	MPI-IO
DDN IME	MPI-IO

	Total I/O size	Scenario
Small	80 GB	Quick data validation
Medium	950 GB	Typical production run
Large	8.4 TB	High-resolution run

TORTIA on pre-GA DDN IME
Total execution time

ICHEC
Irish Centre for High-End Computing

- **6 nodes**
  - 2 x MPI rank /node
  - 20 x OpenMP thread /rank
- **I/O target**
  - In memory
  - Lustre
  - IME Burst Buffer

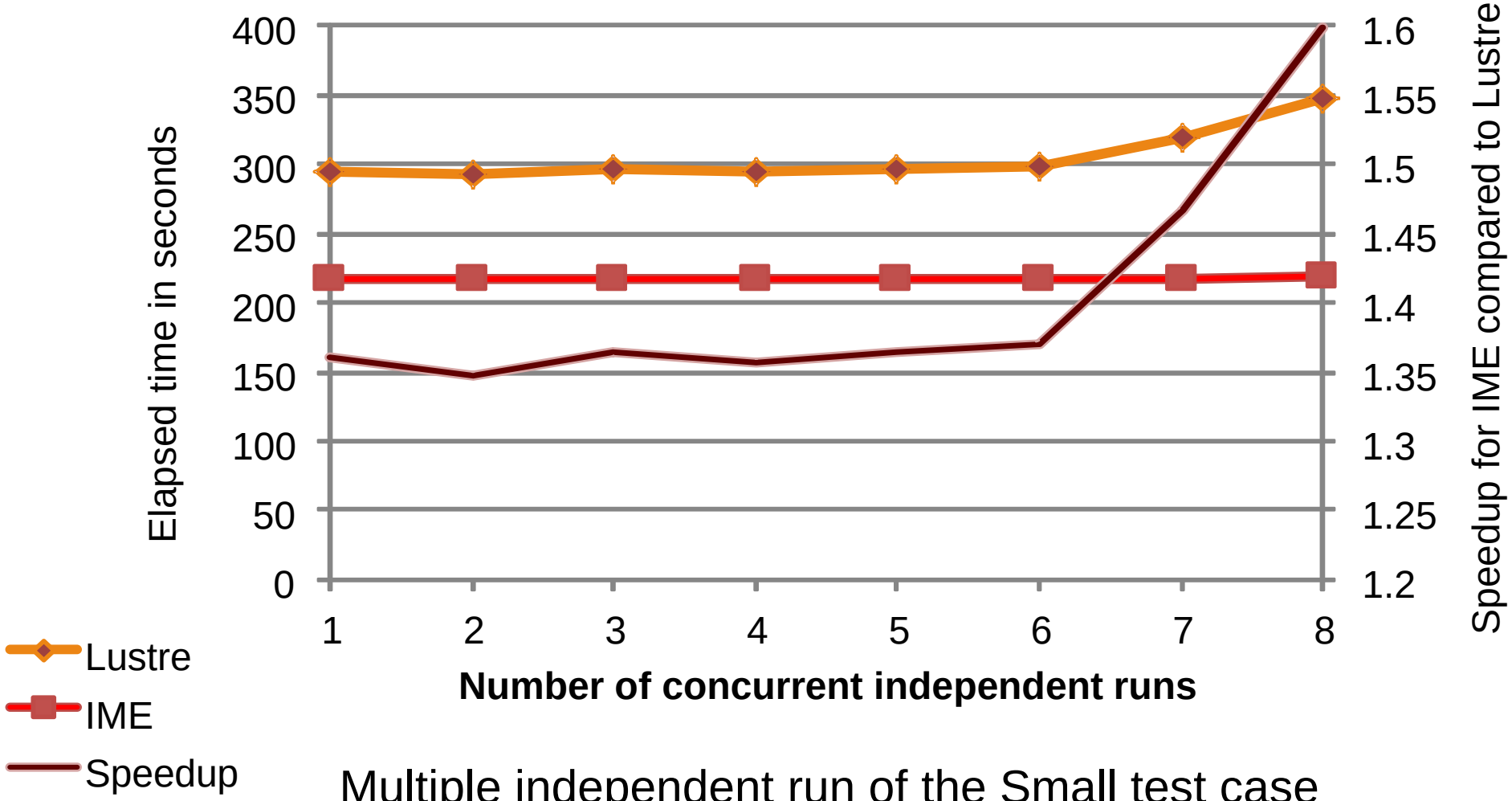


Up-to 3x speedup
Total execution time

In memory not applicable to Large case: not enough memory on the nodes

TORTIA on pre-GA DDN IME Independent run

ICHEC
Irish Centre for High-End Computing



Multiple independent run of the Small test case
1 run x compute node; node count in {1..8}

TORTIA on pre-GA DDN IME
Time spent in I/O

ICHEC
Irish Centre for High-End Computing