Flash-Native Caching for predictable job completion in data intensive environments

Cray User Group - CUG 2017

May 9th 2017

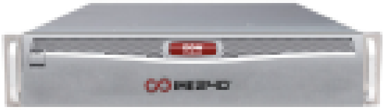
Seattle, WA

Carlos A. A. Thomaz – Technical Product Manager

DDN | IME

DDN’s approach on Burst Buffer and beyond

- **Radical Shift in Performance/Watt,RU,Device**
- **Dramatic Random IO and Shared file IO performance**
- **Self-Optimising in Noisy Environments**
- **Intelligent Read-ahead**
- **Flash-Native implementation**
- **Extreme Rebuild Speeds**
- **Full Data Protection**
- **Improved efficiency of the Parallel Filesystem**

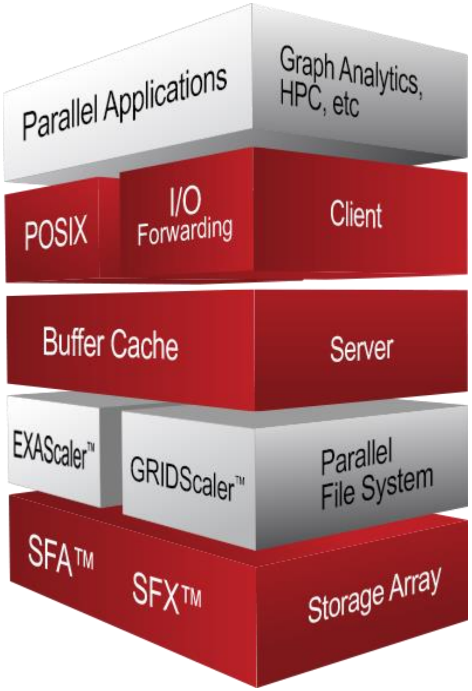


IME240™

DDN | IME

The *Infinite Memory Engine*

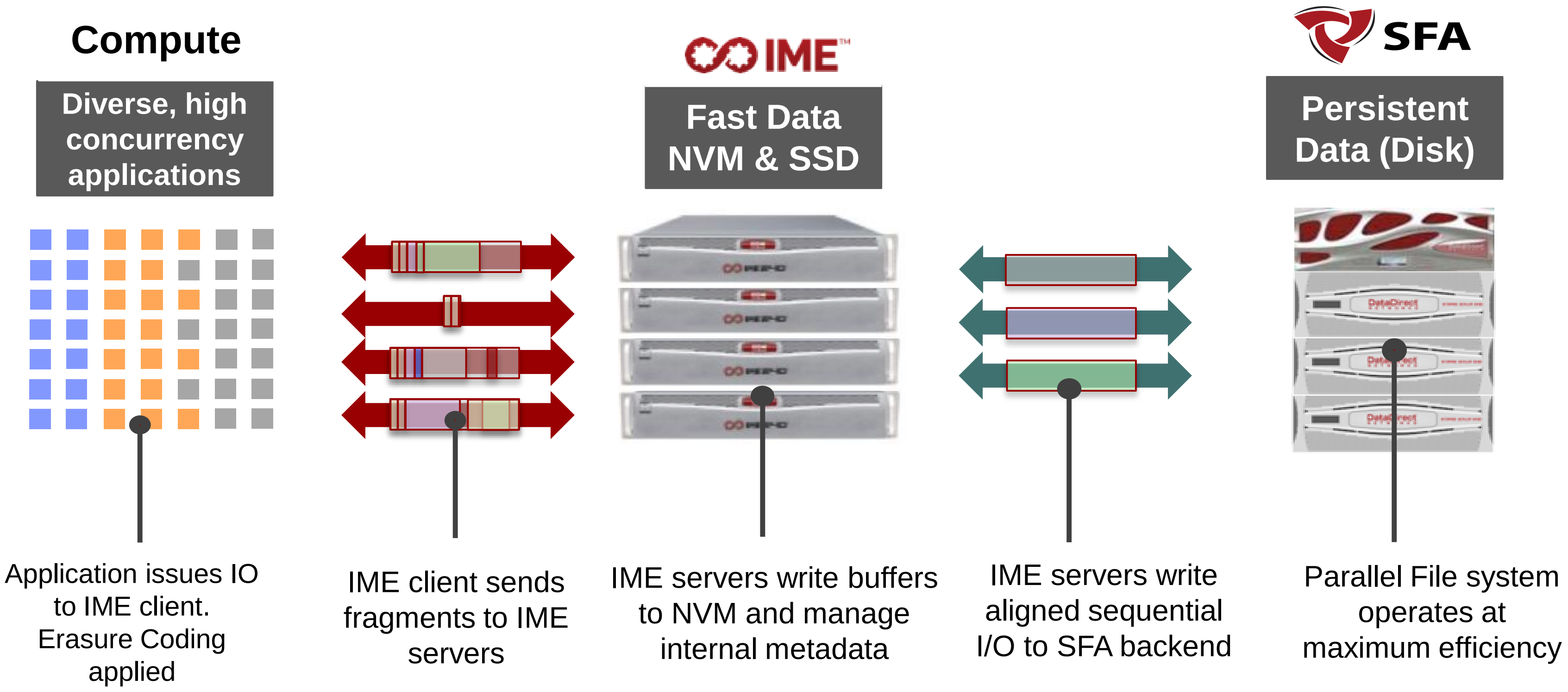
- **A S/W Application Accelerator which leverages NVMe and SSD to remove system level performance bottlenecks**
  - High bandwidth
  - Low latency (Read & Write, Large & Small, Aligned & Random)
  - Data integrity & protection
  - Massive scalability
  - No application changes required



1. *POSIX compatibility for Commercial Big Data Applications*
2. *Solid-state cache provides line-speed performance under almost any I/O profile*
3. *Re-aligns I/O greatly increasing file system performance*
4. *API for job scheduler & application integration*

DDN | IME

I/O dataflow in a nutshell



Compute

Diverse, high concurrency applications

Fast Data NVM & SSD

Persistent Data (Disk)

Application issues IO to IME client. Erasure Coding applied

IME client sends fragments to IME servers

IME servers write buffers to NVM and manage internal metadata

IME servers write aligned sequential I/O to SFA backend

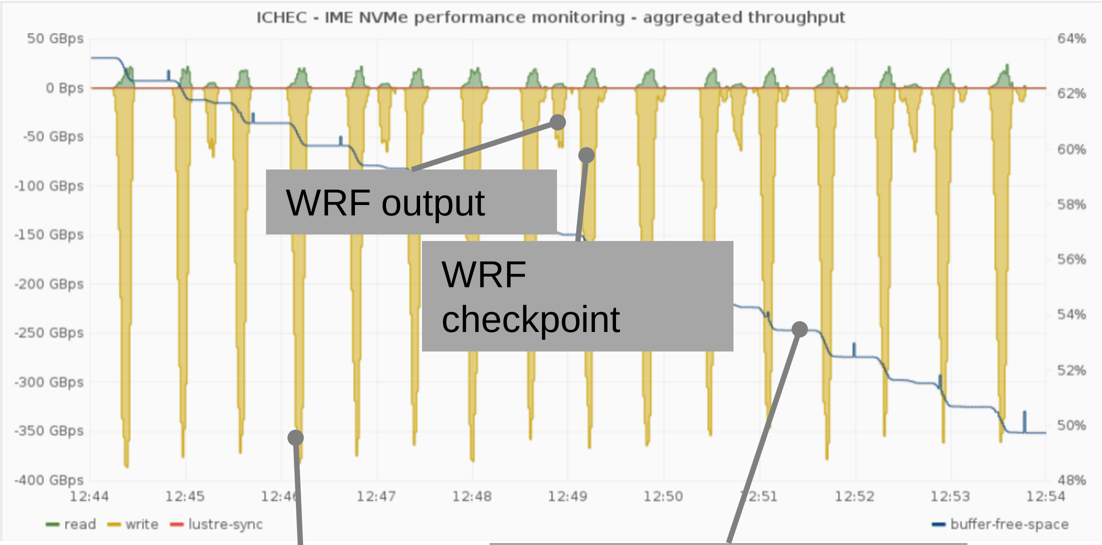
Parallel File system operates at maximum efficiency

WRF | IME

48 jobs across 240 compute nodes

ICHEC
Irish Centre for High-End Computing

48 concurrent MPI job
5 node/job
20 MPI rank/node
IME erasure coding 7+1

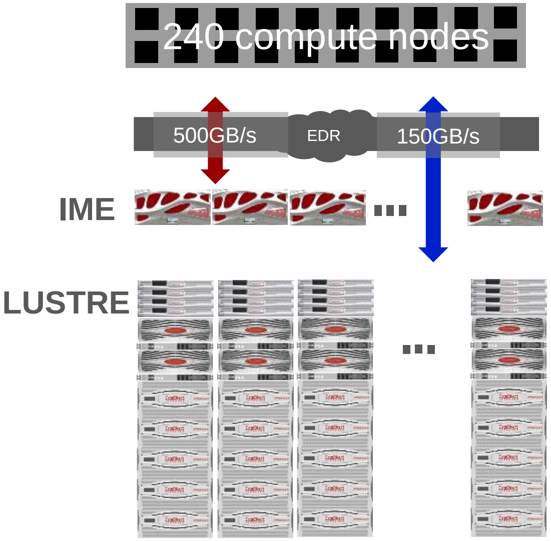


WRF output

WRF checkpoint

~390GB/s

IME capacity consumption



240 compute nodes

500GB/s

EDR

150GB/s

IME

LUSTRE

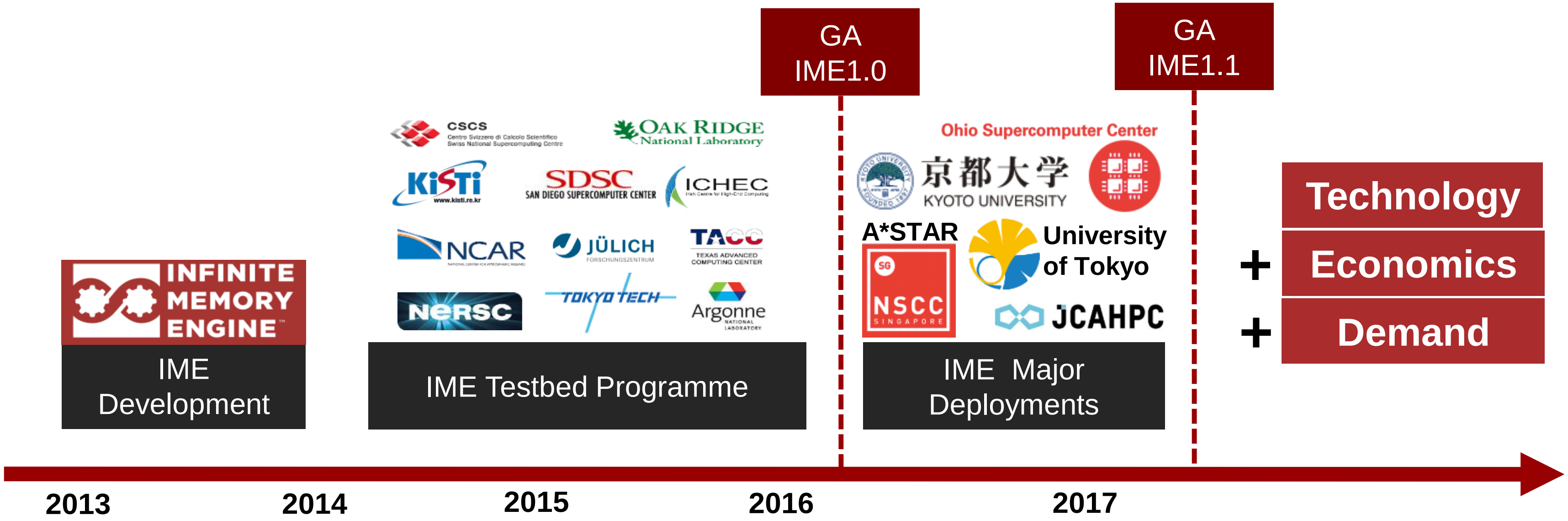
WRF at Scale | IME

Summary Results

	IME		Parallel File system		
	#	Throughput per Metric (GB/s/<x>)	#	Throughput per Metric (GB/s/<x>)	**IME Improvement**
Application Throughput (GB/s)	380		100		**x 3.8**
Rack Units	36	10.5	224	0.45	**x 23**
# IO Nodes	18	21	42	2.4	**x 8.7**
# Drives	432	0.9	2800	0.04	**x 22**
Power Consumption (KW)	27	14	70	1.4	**x 10**

DDN | IME

Timeline



GA IME1.0

GA IME1.1

IME Development

IME Testbed Programme

IME Major Deployments

Technology

+ Economics

+ Demand

2013

2014

2015

2016

2017

INFINITE MEMORY ENGINE
CSCS Centro Svizzero di Calcolo Scientifico Swiss National Supercomputing Centre
OAK RIDGE National Laboratory
KiSTi www.kisti.re.kr
SDSC SAN DIEGO SUPERCOMPUTER CENTER
ICHEC
NCAR
JÜLICH FORSCHUNGSZENTRUM
TACC TEXAS ADVANCED COMPUTING CENTER
NERSC
TOKYO TECH
Argonne NATIONAL LABORATORY
Ohio Supercomputer Center
京都大学 KYOTO UNIVERSITY
A*STAR
NSCC SINGAPORE
University of Tokyo
JCAHPC

Thank You

Interesting to hear more? Find us outside at DDN table

CUG 2017