Next Generation Storage Architectures for Exascale

Mark Seager
CTO for the HPC Ecosystem
Intel Technical Computing Group
Memory Technologies Latency

Cycle time (Rd/Wr)

Memory Decoder ring
SRAM = Static RAM
DRAM = Dynamic RAM
Flash = NAND Flash
PCM = Phase Change Memory
STT = Spin Torque Transfer
FeR = Fero-electric RAM
MRAM = Magnetic RAM
RRAM = Resistance Shift RAM

Highest Performance
Read/Write Energy

R/W Energy per bit (Lower is better)

Memory Decoder ring
SRAM = Static RAM
DRAM = Dynamic RAM
Flash = NAND Flash
PCM = Phase Change Memory
STT = Spin Torque Transfer
FeR = Fero-electric RAM
MRAM = Magnetic RAM
RRAM = Resistance Shift RAM

Lowest energy given higher capacity

SRAM  DRAM  Flash  PCM  STT  FeR  Mram  Rram

nJ

pJ

pJ

0.1

1

10

100

1000

10000

100000
## Memory Technology Score Card

<table>
<thead>
<tr>
<th></th>
<th>SRAM</th>
<th>DRA M</th>
<th>Flash</th>
<th>PCM</th>
<th>STT</th>
<th>FeR</th>
<th>Mram</th>
<th>Rram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>Low</td>
<td>High</td>
<td>V High</td>
<td>V High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Performance</td>
<td>High</td>
<td>Med</td>
<td>Low</td>
<td>Low</td>
<td>Med</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Energy</td>
<td>Low</td>
<td>Med</td>
<td>High</td>
<td>High</td>
<td>Med</td>
<td>?</td>
<td>High</td>
<td>?</td>
</tr>
<tr>
<td>Endurance</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
<td>Med</td>
<td>High</td>
<td>Med</td>
</tr>
<tr>
<td>NV</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Scalable</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Limited</td>
<td>Limited</td>
<td>Limited</td>
</tr>
<tr>
<td>Maturity</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Med</td>
<td>Low</td>
<td>High</td>
<td>Med</td>
<td>Low</td>
</tr>
</tbody>
</table>

- **Small and fast** balanced (capacity, speed, energy)
- High capacity with less activity
New Memory Technologies will Drive a Rethink of Hierarchal Storage Management

<table>
<thead>
<tr>
<th>Node</th>
<th>Cluster</th>
<th>Data Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor/Cache</td>
<td>Processor/Cache</td>
<td>Processor/Cache</td>
</tr>
<tr>
<td>Mem Bus</td>
<td>Mem Bus</td>
<td>Mem Bus</td>
</tr>
<tr>
<td>Memory</td>
<td>Memory</td>
<td>Memory</td>
</tr>
<tr>
<td>I/O Bus</td>
<td>I/O Bus</td>
<td>I/O Bus</td>
</tr>
<tr>
<td>Disk</td>
<td>Fabric</td>
<td>Fabric</td>
</tr>
<tr>
<td></td>
<td>Virt BD</td>
<td>GW</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SAN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Virt BD</td>
</tr>
</tbody>
</table>

- **Node**: 10x BW, 100x IOPS
- **Cluster**: 10x BW, 10x IOPS
- **Data Center**: 10³x BW, 10⁶x IOPS
A new leading edge storage mechanism is required for Exascale

- Design with system focus that enables end-user applications
- Scalable hardware
  - Simple, Hierarchal
  - New storage hierarchy with NVRAM
- Scalable Software
  - Factor and solve
  - Hierarchal with function shipping
- Scalable Apps
  - Asynchronous coms and IO
  - In-situ, in-transit and post processing/visualization
HPC Software that Exascales up and also scales down for transparent user experience

- New hierarchal resilient distributed application object storage model
- Burst buffer
- SMP scalability
- Namespace [DNE] scalability
- Improved resilience
- New back-end fs

Storage mechanisms are a critical system component

### Applications
- MKL
- Fortran
- Python
- TV, ADDT "SMPA"
- UPC
- GNU, Intel
- MRNet
- Lustre
- SLURM
- Shell
- SLURM Utils
- (Distro)

### Runtime libraries (Distro)
- Intel
- GNU

### Compute node: CNOS
- Service node: Linux Kernel (Distro)

### RAS
- Platform SW (not firmware)
New approach to storage hierarchy: applications driven object oriented data storage

- UQ, Applications define objects
- Storage of objects is abstracted
- Includes remote method invocation for user computations near the data
- Access transformed from shell+ls ➞ Python
- Metadata is accreted during object creation and IO
- Enables distributed data intensive computing model
- Enables Lustre ecosystem
- Enables analytics
The “Data Challenge”

“Every two days, we create as much information as we did from the dawn of civilization up until 2003.”

— Eric Schmidt, former Google CEO

... and this is only the beginning
Big Data Graphs are Everywhere

Over 24 Petabytes
Data processed by Google every day in 2011

7 Exabytes
Data traffic by mobile users worldwide in 2011

4 billion
Pieces of content shared on Facebook every day by July 2011

250 Million
Tweets per day in Oct 2011

5.5 million
Legitimate emails sent every second in 2011

158 products
Ordered per second on Cyber Monday in 2010

1500+ blog posts
Every minute in 2011

Internet devices: 1000 billion by 2013
Up from 5 billion in 2010

Internet traffic to increase 9x by 2013
From 5 Exabyte a month to 56 Exabyte a month in 2013

More video was uploaded to YouTube
In last 2 months, than if ABC, NBC, and CBS had been airing new content since 1948

Between the birth of the world and 2003, there were 5 Exabyte of information created. We now create 5 Exabyte every 2 days.

Eric Schmidt

... and graphical analysis is getting more and more sophisticated.
Grand Challenge: Knowledge Extraction

Storage and Traffic growing exponentially... and what’s vacuumed up is processed using Analytics, Machine Learning, and Data Mining methods.

Big Data plays a big role in the Cloud

- 24 Million Wikipedia Pages
- 750 Million Facebook Users
- 6 Billion Flickr Photos
- 48 Hours a Minute YouTube

Growing faster than Moore’s Law
Optimized Storage Enables...

DATA  INFORMATION  INSIGHT

"Insight" – the Ultimate Goal
Doing this at commercial scale...

... requires some form of distributed computation.

Data-Parallel  Graph-Parallel

Cluster Computing Architecture  GraphLab

What I Like  What My Friends Like
Intel portfolio delivers balanced performance

Shown to improve 1 Terabyte sort from 4 hours to 7 minutes
>34x improvement

- Intel® Xeon® E5-2690 processor ~50% improved
- Intel® SSD 520 Series ~80% improved
- Intel® 10GbE Adapters ~50% improved
- Intel® Distribution for Apache Hadoop® software ~40% improved

Other brands and names are the property of their respective owners.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Source: Intel Internal testing
For more information go to: intel.com/performance

~40% improved
~50% improved
~80% improved
~50% improved
~7 minutes
~7 minutes
>4 hours
>4 hours
Leveraging DAOS into Big Data: Arbitrarily Connected Graph Data Analytics

- Many large-scale machine learning problems involve graph structures, and **Hadoop** is ideal for constructing graphs for Exascale computations:
  - Graph relationships built from unstructured data
  - Objects/relationships stored to DAOS via self-describing data API (HDF5) and then loaded by Exascale
- **GraphLab’s** asynchronous execution model is ideal for a wide range of machine learning computations
  - Each node processes a portion of graph
  - Objects loaded from HDF5/DAOS during execution as needed
- After computation, DAOS may be used by various cloud services to query selected object values
- Intel Lab’s prototyping effort:
  - Port Hadoop and GraphLab to the new DAOS interface
  - Evaluate functionality on COTS systems
  - Evaluate ingress and execution performance on Exascale prototype using large-scale machine learning benchmarks

**DAOS will serve as the bridge between multiple big data paradigms and also HPC**
FastForward funded Big Data – HPC Bridge Architecture

**ACG Ingress on a Hadoop Cluster**
- Node
- Node
- Node
- Node

ACG Ingress Processing
HDF5 Adaptation Layer
HDF5

**HPC**
- Node
- Node
- Node
- Node

Computation Kernel
HDF5 Adaptation Layer
HDF5

Graph (Partitions) and Network Information Represented in HDF5/DAOS

Raw Data

Big Data – HPC Bridge

Results
The next generation storage paradigm spans HPC and BigData

- **Conventional namespace**
  - Works at human scale
  - Administration, security, accounting
  - Supports legacy data and applications

- **DAOS Containers**
  - Work at exascale
  - Separate scalable object namespace
    - Application data + metadata
    - High-level I/O models determine schema
    - Object reference invariance
  - Transactional

- **Storage pools**
  - Administer by usage
    - Small / random, Large / streaming
    - Low capacity / fast, High capacity / slow
  - Migration
New memory technologies and Exascale drive a different HPC storage paradigm.

Big Data and HPC have similar requirements.

Let's go fully object oriented. The time is now!
Legal Information

Today's presentations contain forward-looking statements. All statements made that are not historical facts are subject to a number of risks and uncertainties, and actual results may differ materially. Please refer to our most recent Earnings Release and our most recent Form 10-Q or 10-K filing for more information on the risk factors that could cause actual results to differ.

If we use any non-GAAP financial measures during the presentations, you will find on our website, intc.com, the required reconciliation to the most directly comparable GAAP financial measure.

INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS”. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, reference www.intel.com/software/products.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
Legal Disclaimers

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. Go to: http://www.intel.com/products/processor_number

Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM). Functionality, performance or other benefits will vary depending on hardware and software configurations. Software applications may not be compatible with all operating systems. Consult your PC manufacturer. For more information, visit http://www.intel.com/go/virtualization

No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology (Intel® TXT) requires a computer system with Intel® Virtualization Technology, an Intel TXT-enabled processor, chipset, BIOS, Authenticated Code Modules and an Intel TXT-compatible measured launched environment (MLE). Intel TXT also requires the system to contain a TPM v1.s. For more information, visit http://www.intel.com/technology/security

Requires a system with Intel® Turbo Boost Technology. Intel Turbo Boost Technology and Intel Turbo Boost Technology 2.0 are only available on select Intel® processors. Consult your PC manufacturer. Performance varies depending on hardware, software, and system configuration. For more information, visit http://www.intel.com/go/turbo

Intel® AES-NI requires a computer system with an AES-NI enabled processor, as well as non-Intel software to execute the instructions in the correct sequence. AES-NI is available on select Intel® processors. For availability, consult your reseller or system manufacturer. For more information, see http://software.intel.com/en-us/articles/intel-advanced-encryption-standard-instructions-aes-ni/

Intel, Intel Xeon, the Intel Xeon logo and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. Other names and brands may be claimed as the property of others.

Copyright © 2012, Intel Corporation. All rights reserved.