SSD/NVM Technology Boosting Ceph Performance

JACK Zhang, Enterprise Architect, Intel

Xinxin Shu, Software Engineer, Intel

Ping She, Technical Mkt Engineer, Intel

10/2015

Agenda - Overview

- SSDs for Ceph today, The future SSDs is here -NVM Express™
- 3D NAND and 3D XPoint™ for Ceph tomorrow
- Yahoo! Case study w/Intel NVMe SSD+Intel Cache Acceleration software
- ALL SSD Ceph performance data reviews
- Summary, Q&A

Agenda

- SSDs for Ceph today, The future SSDs is here -NVM Express™
- 3D NAND and 3D XPoint™ for Ceph tomorrow
- Yahoo! Case study w/Intel NVMe SSD+Intel Cache Acceleration software
- ALL SSD Ceph performance data reviews
- Summary, Q&A

Typical SSD usage today at Ceph

SSD as Journal and Caching

- SSD as Journal drive + Caching drive
- Intel Cache Acceleration Software (CAS with hinting technology is optimized for Ceph

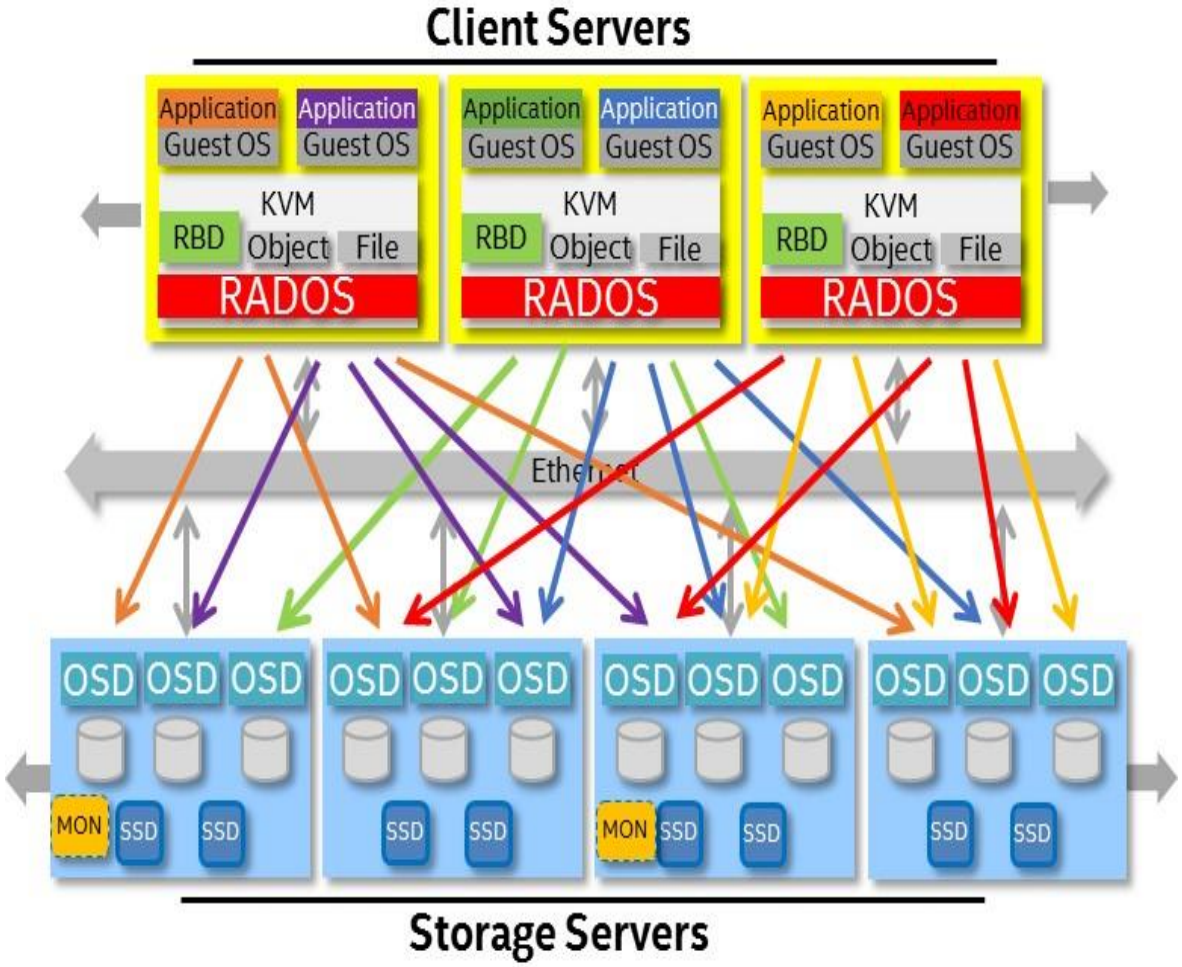
  Example, Intel CAS + Intel P3700 for Yahoo! Ceph
- Separate Journal and Caching with different SSD

SSD : HDD ratio (recommendation)

- Not all SSDs are same, different density may have different performance
- Example, Intel SSD as Journal

  SATA SSD: 1 S3700: 5 HDDs

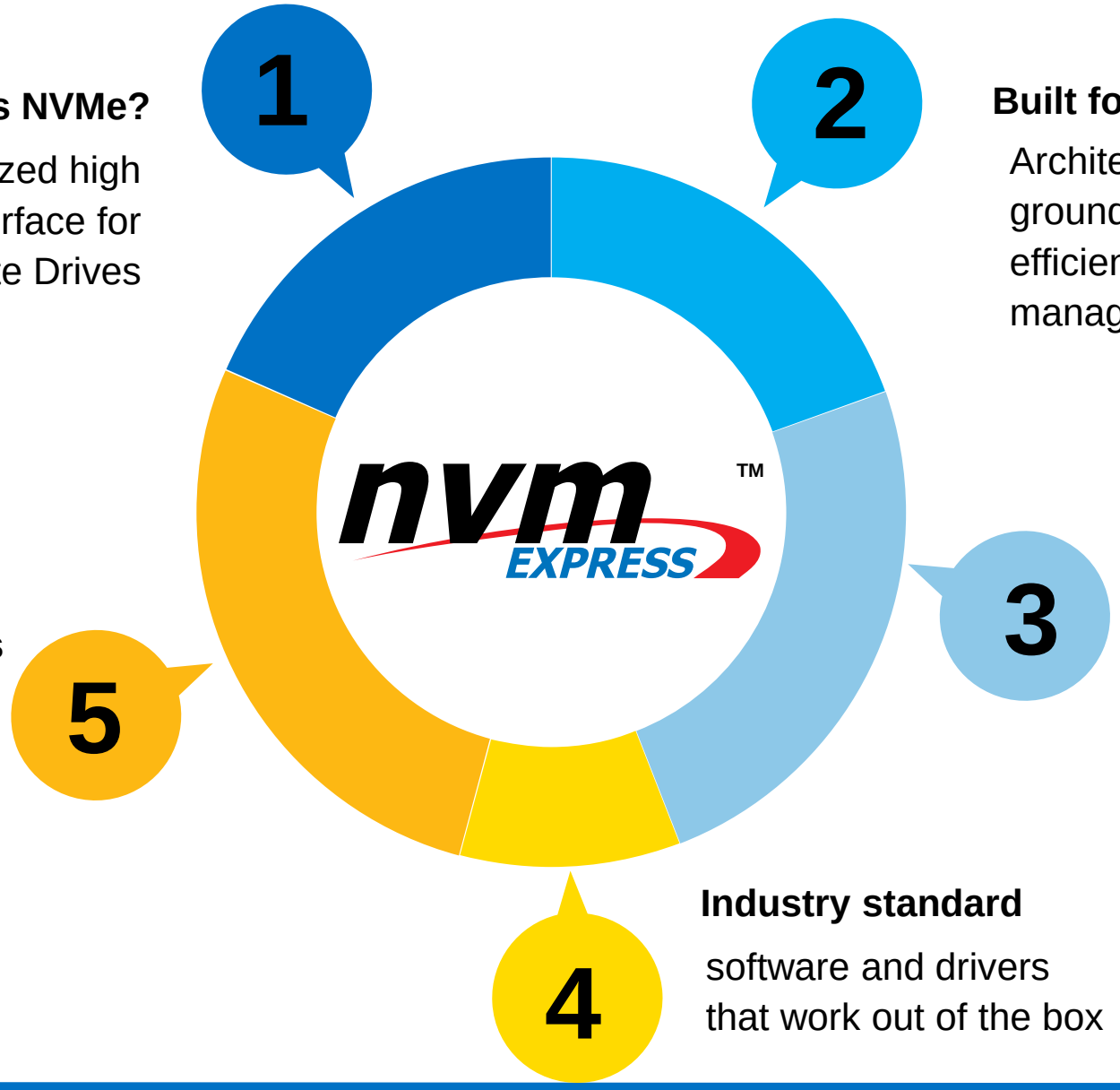
  PCIe/NVMe SSD: 1 P3700 : 20 HDDs



The Future is Here – NVM Express™

What is NVMe?

NVM Express is a standardized high performance software interface for PCI Express® Solid-State Drives



Built for SSDs

Architected from the ground up for SSDs to be efficient, scalable, and manageable

Developed to be lean

Streamlined protocol with new efficient queuing mechanism to scale for multi-core CPUs, low clock cycles per IO

Industry standard

software and drivers that work out of the box

Ready for next generation SSDs

New storage stack with low latency and small overhead to take full advantage of next generation NVM

ALL SSD Solutions for Ceph

- All NVMe/PCIe SSDs

  Best performance but:

  a) High cost, low capacity (today)

  b) NIC bandwidth

- NVMe + Low Cost SATA high density SSDs

  a) Best TCO for performance, can have higher storage density

  b) Example, P3700 800GB for Journal drive, 4x S3510 1.6TB as data

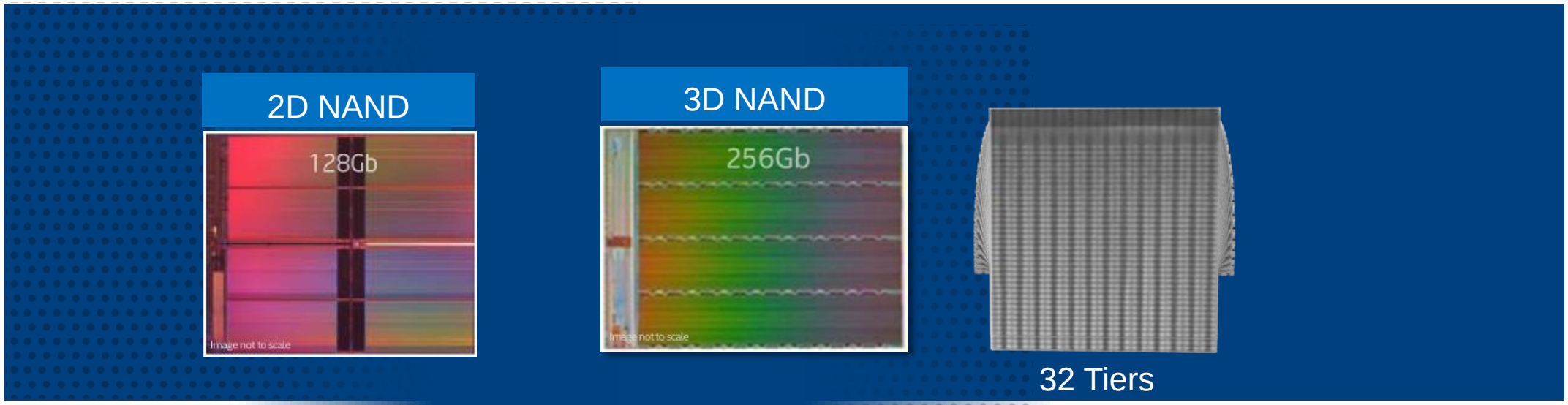
- SSDs for Client nodes

Agenda

- SSDs for Ceph today, The future SSDs is here -NVM Express™
- 3D NAND and 3D XPoint™ for Ceph tomorrow
- Case study, Intel NVMe SSD+iCAS for Yahoo Ceph
- ALL SSD Ceph performance data reviews
- Summary, Q&A

Intel Continues to Drive Technology

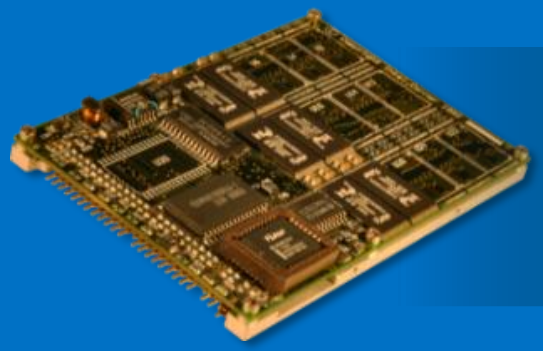
Accelerating Solid State Storage in Computing Platforms



CAPACITY	Enables high-density flash devices
COST	Achieves lower cost per gigabyte than 2D NAND at maturity
CONFIDENCE	3D architecture increases performance and endurance

Moore’s Law Continues to Disrupt the Computing Industry

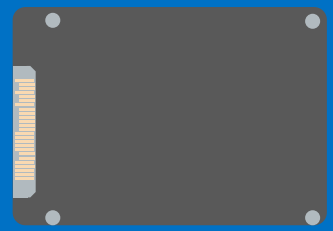
1992 12MB



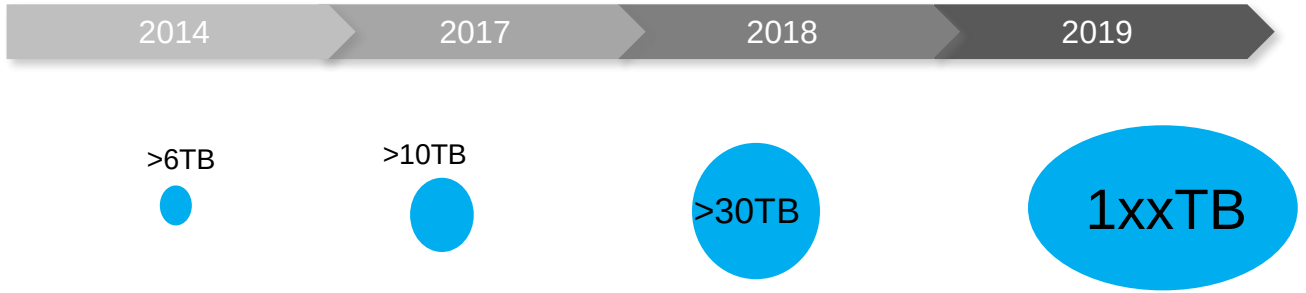
First Intel® SSD for Commercial Usage

1,000,000x
the capacity while shrinking the form factor

2017 >10TB



U.2 SSD



Source: Intel projections on SSD capacity

3D XPoint™ TECHNOLOGY

A new class of non-volatile memory Media

1000X
FASTER
THAN NAND[1]

1000X
ENDURANCE
OF NAND[1]

10X
DENSER
THAN DRAM[1]

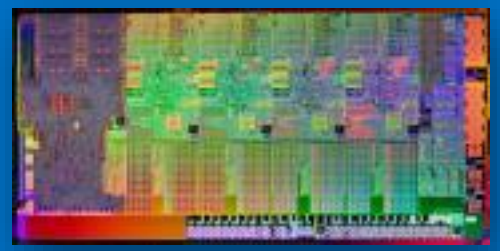
NAND-LIKE DENSITIES AND DRAM-LIKE SPEEDS

[1]Technology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of in-market memory products against internal Intel specifications

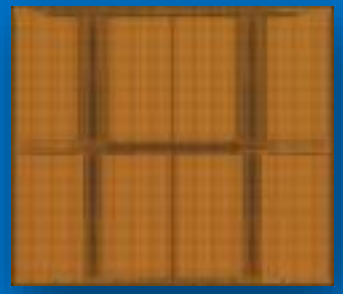
3D XPoint™ TECHNOLOGY

Breaks the Memory Storage Barrier

SRAM
Latency: 1X
Size of Data: 1X



DRAM
Latency: ~10X
Size of Data: ~100X



STORAGE

3D XPoint ™ Memory Media
Latency: ~100X
Size of Data: ~1,000X

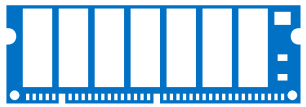
NAND SSD
Latency: ~100,000X
Size of Data: ~1,000X

HDD
Latency: ~10 MillionX
Size of Data: ~10,000X

MEMORY

Technology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of in-market memory products against internal Intel specificatio

New Faster SSDs Emerge, Where Will They be Used?



Extend DRAM

In memory database
Key value store, memcache, RDMA replacement, NEW Applications/Businesses
goal is to lower TCO and execute larger datasets

Faster SSD

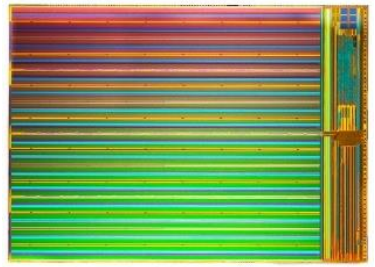
Real time analytics – targeting noSQL databases
Fraud detection, ad bidding, real time decision making, trading
Cloud hosting – amazing QoS for better application SLAs
Ultra High Definition all professional video production

Caching

Cloud hosting, all flash array, SAN
Write buffer for RAID array

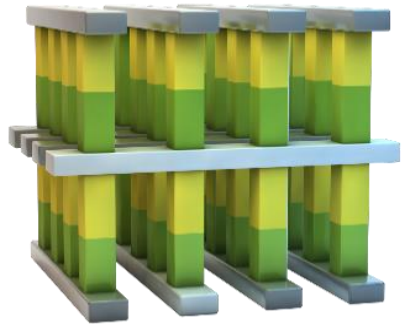
NAND Flash vs 3D XPoint™ Technology for Ceph tomorrow

3D MLC and TLC NAND



Enable higher capacity OSDs at lower price

3D XPoint™ Technology



Higher performance, opening up new use cases, DRAM extended, Key/value…

Agenda

- SSDs for Ceph today, The future SSDs is here -NVM Express™
- 3D NAND and 3D XPoint™ for Ceph tomorrow
- Yahoo! Case study w/Intel NVMe SSD+Intel Cache Acceleration software
- ALL SSD Ceph performance data reviews
- Summary, Q&A

Intel® NVMe SSD Accelerates Ceph for YAHOO!

CHALLENGE:

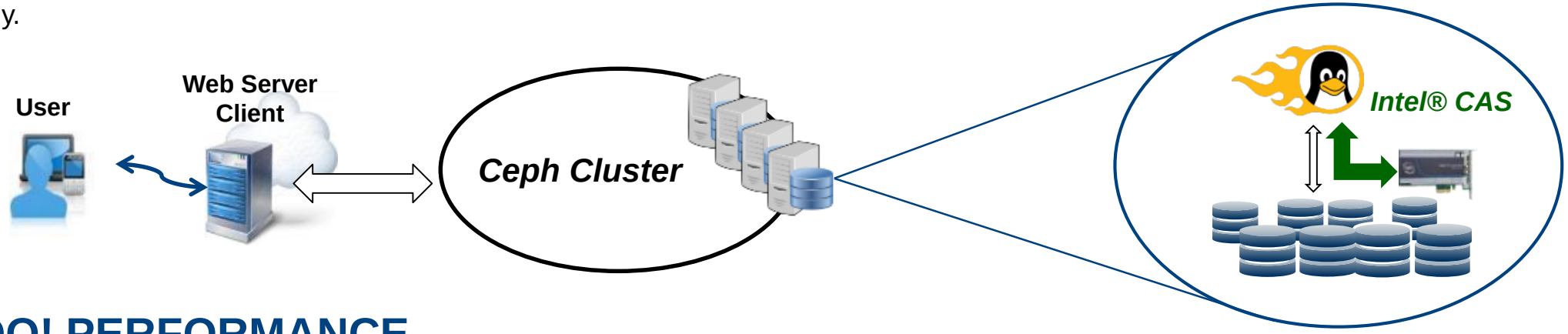
- How can I use Ceph for scalable storage solution?

(High latency and low throughput due to erasure coding, write twice, huge number of small files)

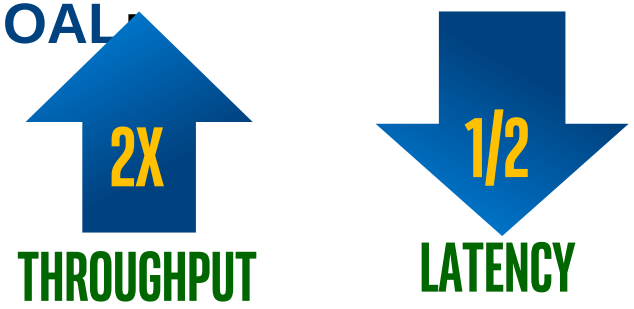
- Use over-provision to address performance is costly.

SOLUTION:

- Intel® NVMe SSD – consistently amazing
- Intel® CAS 3.0 feature – hinting
- Intel® CAS 3.0 fine tuned for Yahoo! – cache metadata



YAHOO! PERFORMANCE GOAL:



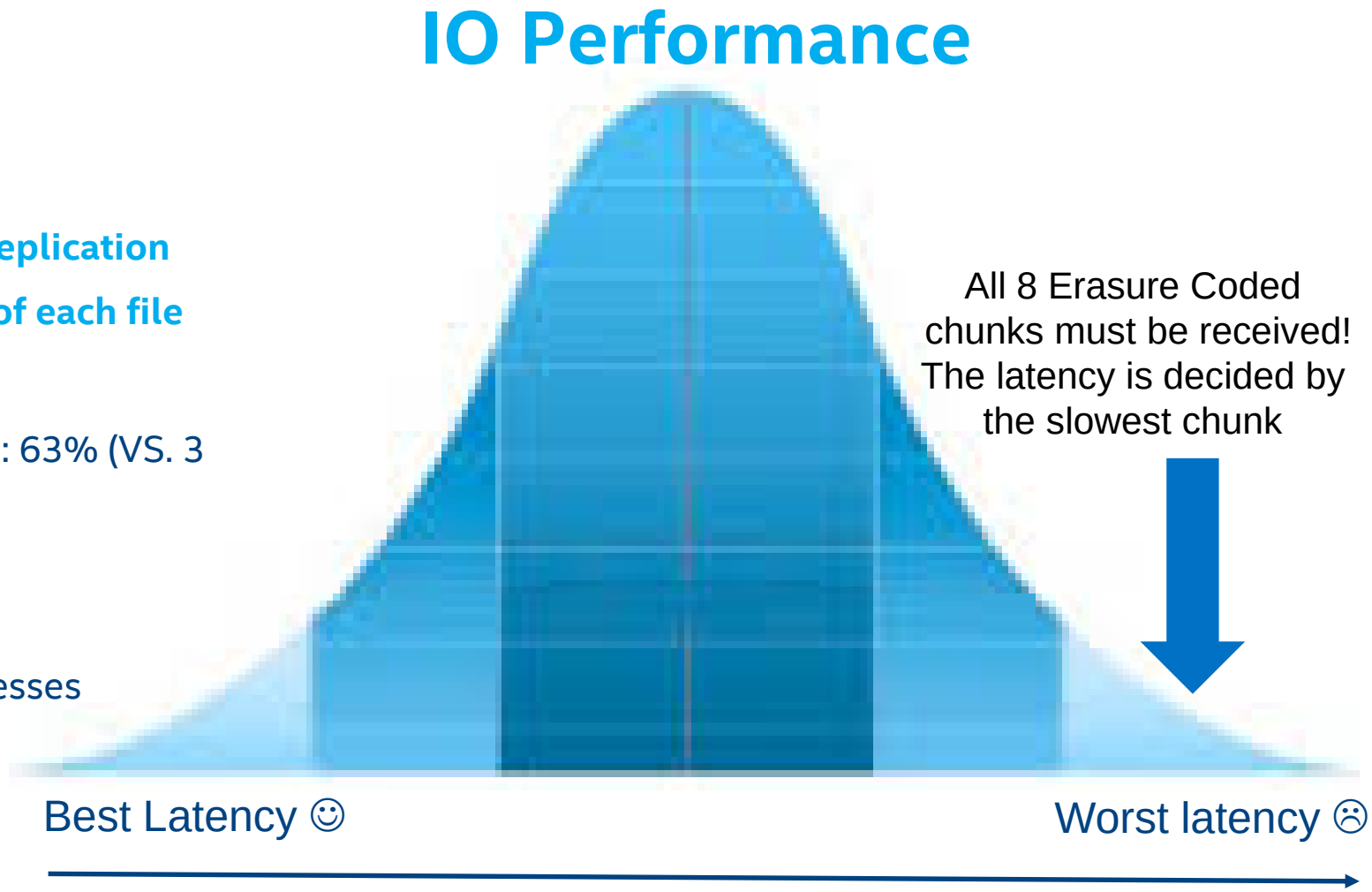
COST REDUCTION:

- ***CapEx savings*** *(over-provision ↓)*
  OpEx savings *(Power, Space, Cooling… ↓)*
- ***Improved scalability planning*** *(Performance and Predictability ↑)*

Yahoo! CEPH Challenges

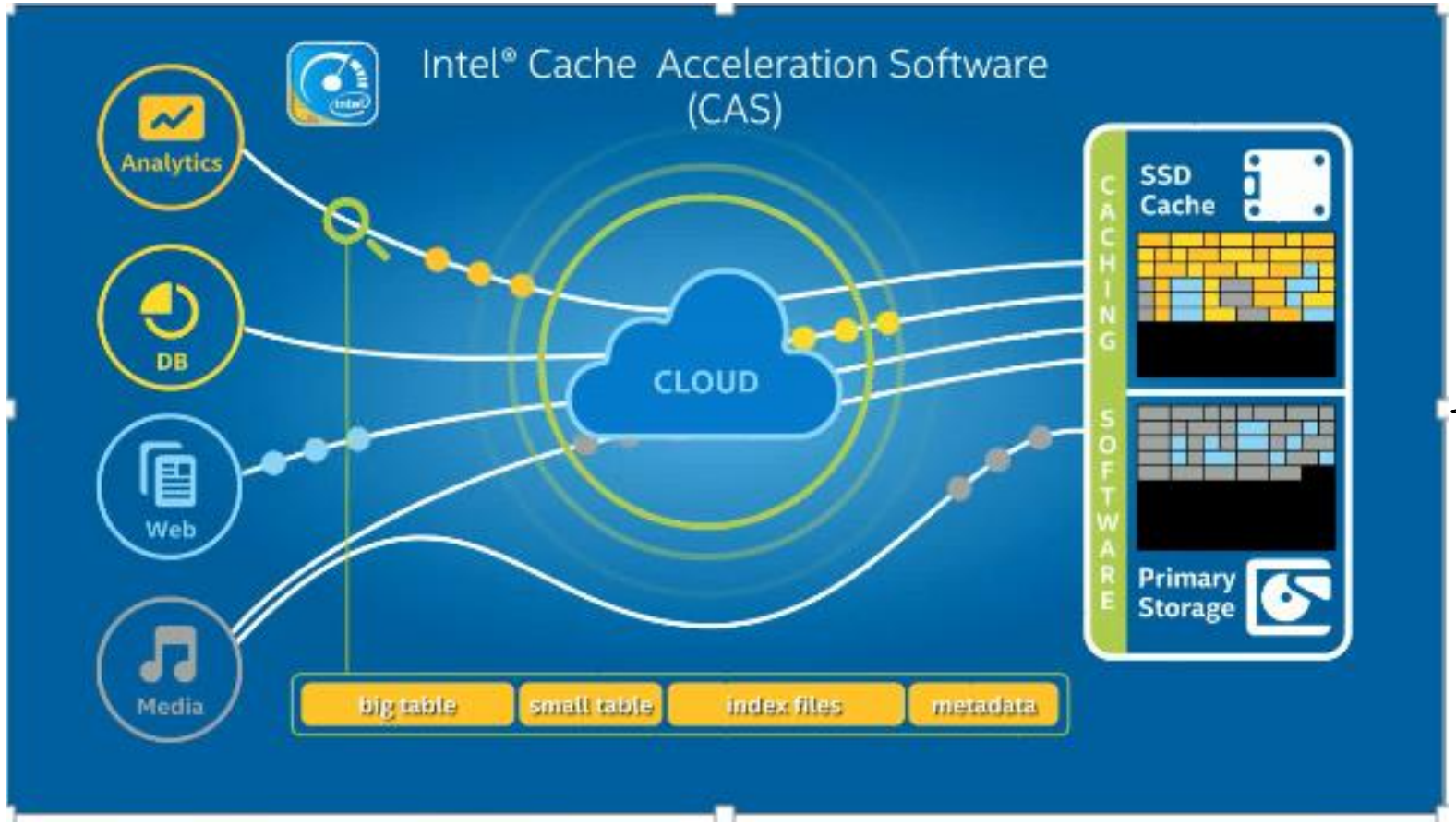
Huge number of small files, why?

- **Write twice**
- **Erasure Code**
  - **Great for data efficiency –vs- RAID1 replication**
  - **But bad for # of file and smaller size of each file**
- Example:
  - Erasure Coding (8+3) higher utilization: 63% (VS. 3 replications: 25%)
  - 1M photo: become 11 x 128K
  - Number of files: 64 – 128 millions files
  - One file access: become 3-4 disks accesses



Intel CAS Linux 3.0 Feature – Hinting

Not all data equal. As cache, we treat them differently.



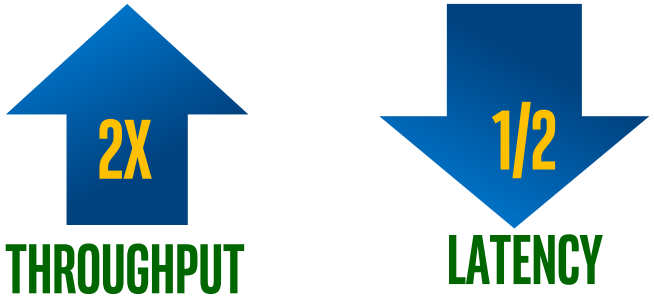
Click to play video

Compelling Solution. Now What?

- ***Consideration to adopt***
  - Use Intel NVMe SSD as cache
  - Intel CAS Linux 3.0 with hinting feature will be released by end of this year
  - Support RHEL, SLES, CentOS, ext4, ext3, xfs.
  - Intel will help to fine tune performance for your Ceph workload
  - Sign up with us as early engagement customer

- ***Take Away Message: 5% caching for 2X performance!***

YAHOO! PERFORMANCE GOAL:

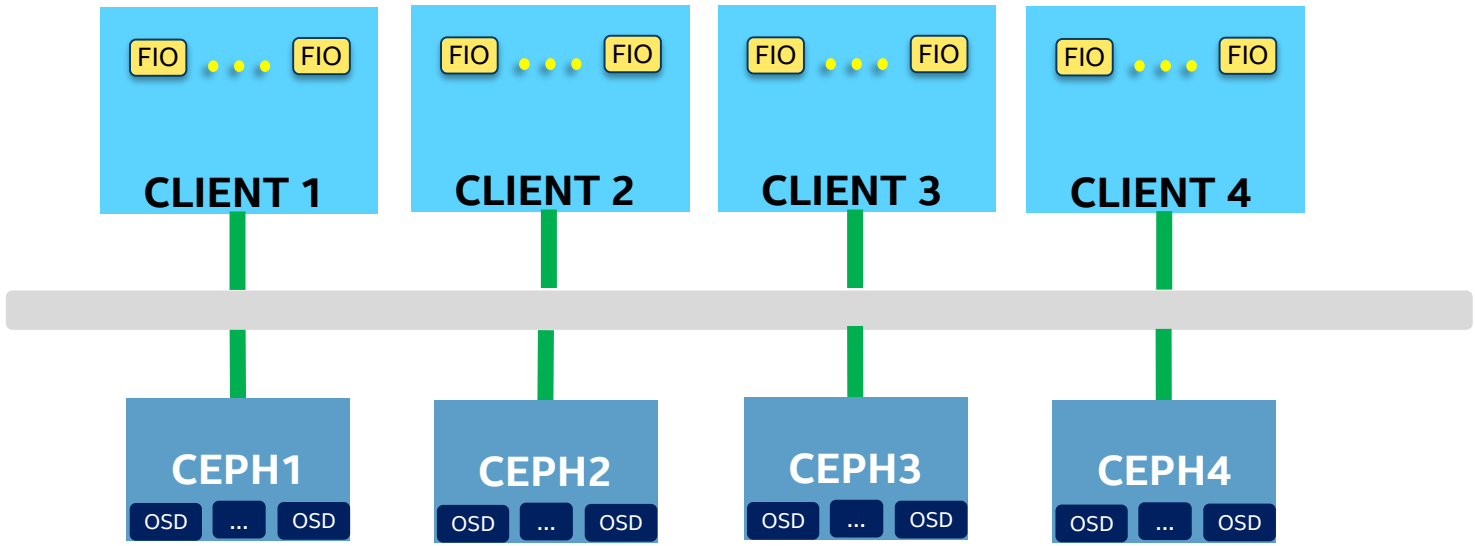


- ***To Learn More***
  - Ceph IDF 2015 Demo:
    https://www.youtube.com/watch?v=vtIIbxO4Zlk
  - Special yahoo speaker IDF 2015:
    http://intelstudios.edgesuite.net//idf/2015/sf/aep/SSDS002/SSDS002.html

Agenda

- SSDs for Ceph today, The future SSDs is here -NVM Express™
- 3D NAND and 3D XPoint™ for Ceph tomorrow
- Yahoo! Case study w/Intel NVMe SSD+Intel Cache Acceleration software
- ALL SSD Ceph performance data reviews
- Summary, Q&A

System Configuration



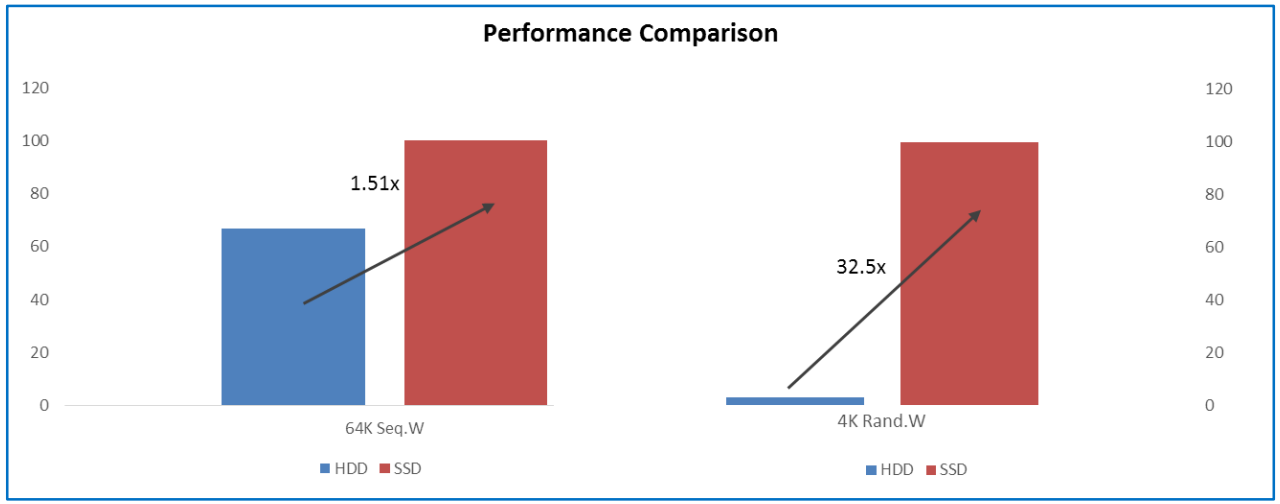
Client Node

- 4 nodes with Intel® Xeon™ processor E5-2699 v3 @ 2.30GHz, 64GB memory
- OS : Ubuntu Trusty

Storage Node

- 4 nodes with Intel® Xeon™ processor E5-2699 v3 @ 2.30GHz, 64GB memory
- Ceph Version : 0.94.2
- OS : Ubuntu Trusty
- SSD Setup
  - 16 DC 3700 400 GB for OSD
  - 4 P3600 SSD for Journal

SSD Cluster vs. HDD Cluster



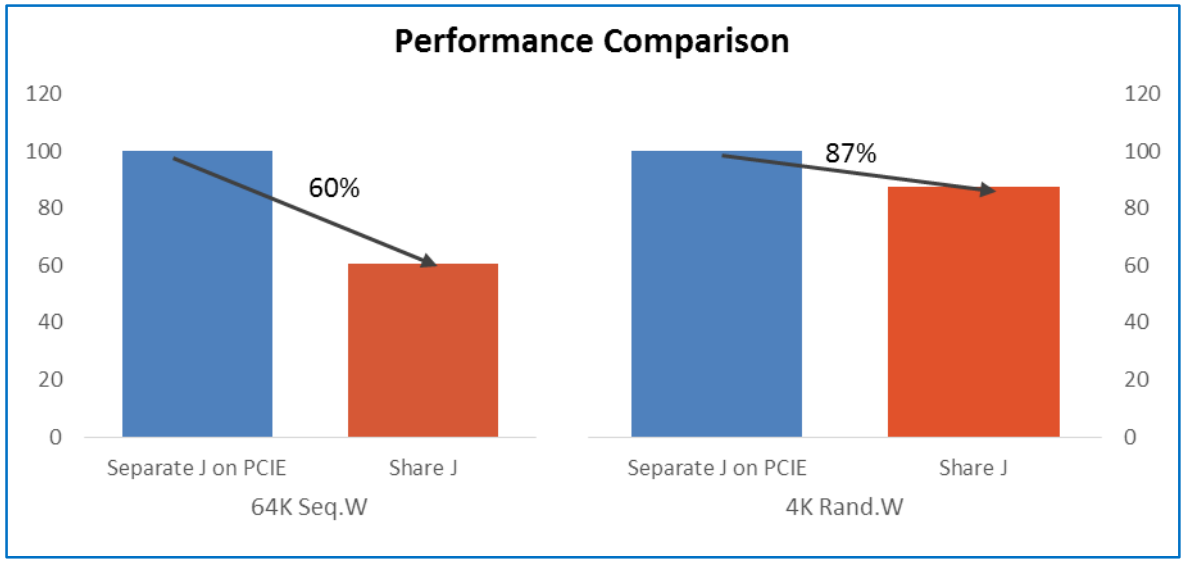
- ***Compared to HDD cluster***
  - 40 HDDs with journal on SSD
  - For 4K random write, need ~ 32x HDD Cluster (~ 1300 HDDs) to get same performance
  - For 64K Sequential write, need ~ 1.5x HDD Cluster (~ 60 HDDs) to get the same performance

For the use cases that require high performance, using SSD can significantly reduce TCO

Comparison with Journal on the same SSD

Deployments

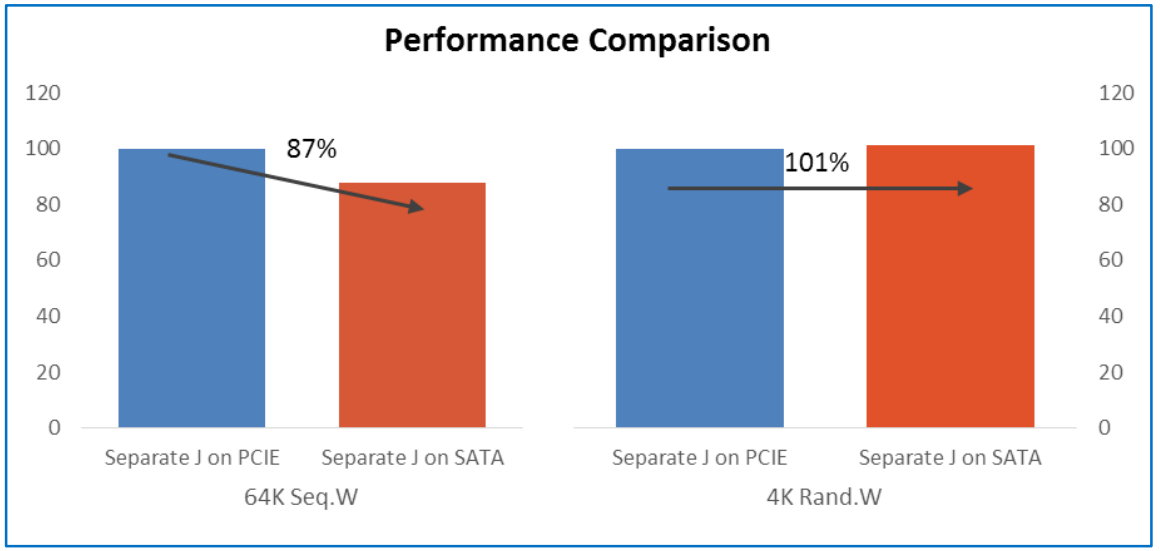
1. **Journal on separate PCIe SSD**
2. **Journal on same SATA SSD**



~67% higher for 64K Sequential write & 14% higher for 4K random write.

Comparison with Journal on the Separate SATA SSD

Deployments
Journal on separate PCIe SSD
Journal on separate SATA SSD



Same IOPS for 4K random write, ~ 14% higher performance for 64K Sequential write. Using NVMe SSD as journal can spare more slots so maybe we can eliminate external storage enclosures

Agenda

- SSDs for Ceph today, The future SSDs is here -NVM Express™
- 3D NAND and 3D XPoint™ for Ceph tomorrow
- Yahoo! Case study w/Intel NVMe SSD+Intel Cache Acceleration software
- ALL SSD Ceph performance data reviews
- Summary, Q&A

Summary

- NVMe™ was built for high-performance SSDs with the future in mind - and ready today, using PCIe SSD as journal can get higher performance and eliminate external storage enclosure
- Intel iCAS with hinting is a leading caching solution
- All SSD solutions provide best ever performance, and bright TCO, while there are still lots of space for further improvements
- 3D NAND is the building block for the high capacity, low cost SSDs –- candidate for future OSD drives
- 3D XPoint™ technology delivers high performance and low latency –- opening new fastest NVM applications/usages

Q & A

Legal Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

The products described 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.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, Xeon and the Intel logo are trademarks of Intel Corporation in the United States and other countries.

*Other names and brands may be claimed as the property of others.

Legal Information: Benchmark and Performance Claims Disclaimers

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.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.

Test and System Configurations: See Back up for details.

For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the first quarter, the year and the future are forward-looking statements that involve a number of risks and uncertainties. Words such as "anticipates," "expects," "intends," "plans," "believes," "seeks," "estimates," "may," "will," "should" and their variations identify forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel's actual results, and variances from Intel's current expectations regarding such factors could cause actual results to differ materially from those expressed in these forward-looking statements. Intel presently considers the following to be important factors that could cause actual results to differ materially from the company's expectations. Demand for Intel's products is highly variable and could differ from expectations due to factors including changes in the business and economic conditions; consumer confidence or income levels; customer acceptance of Intel's and competitors' products; competitive and pricing pressures, including actions taken by competitors; supply constraints and other disruptions affecting customers; changes in customer order patterns including order cancellations; and changes in the level of inventory at customers. Intel's gross margin percentage could vary significantly from expectations based on capacity utilization; variations in inventory valuation, including variations related to the timing of qualifying products for sale; changes in revenue levels; segment product mix; the timing and execution of the manufacturing ramp and associated costs; excess or obsolete inventory; changes in unit costs; defects or disruptions in the supply of materials or resources; and product manufacturing quality/yields. Variations in gross margin may also be caused by the timing of Intel product introductions and related expenses, including marketing expenses, and Intel's ability to respond quickly to technological developments and to introduce new features into existing products, which may result in restructuring and asset impairment charges. Intel's results could be affected by adverse economic, social, political and physical/infrastructure conditions in countries where Intel, its customers or its suppliers operate, including military conflict and other security risks, natural disasters, infrastructure disruptions, health concerns and fluctuations in currency exchange rates. Results may also be affected by the formal or informal imposition by countries of new or revised export and/or import and doing-business regulations, which could be changed without prior notice. Intel operates in highly competitive industries and its operations have high costs that are either fixed or difficult to reduce in the short term. The amount, timing and execution of Intel's stock repurchase program and dividend program could be affected by changes in Intel's priorities for the use of cash, such as operational spending, capital spending, acquisitions, and as a result of changes to Intel's cash flows and changes in tax laws. Product defects or errata (deviations from published specifications) may adversely impact our expenses, revenues and reputation. Intel's results could be affected by litigation or regulatory matters involving intellectual property, stockholder, consumer, antitrust, disclosure and other issues. An unfavorable ruling could include monetary damages or an injunction prohibiting Intel from manufacturing or selling one or more products, precluding particular business practices, impacting Intel's ability to design its products, or requiring other remedies such as compulsory licensing of intellectual property. Intel's results may be affected by the timing of closing of acquisitions, divestitures and other significant transactions. A detailed discussion of these and other factors that could affect Intel's results is included in Intel's SEC filings, including the company's most recent reports on Form 10-Q, Form 10-K and earnings release.