# Ceph at Intel
# Several Examples of Our Work

Dan Ferber
Storage Group, Intel Corporation
SC15 with University of Minnesota
November 2015

# Legal Disclaimer

Notice: This document contains information on products in the design phase of development. The information here is subject to change without notice. Do not finalize a design with this information.

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. Intel does not assume any liability for lost or stolen data or systems or any damages resulting from such losses.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

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

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.

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 Intel product specifications and roadmaps.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

<Use only if applicable> Warning: Altering PC clock or memory frequency and/or voltage may (i) reduce system stability and use life of the system, memory and processor; (ii) cause the processor and other system components to fail; (iii) cause reductions in system performance; (iv) cause additional heat or other damage; and (v) affect system data integrity. Intel assumes no responsibility that the memory, included if used with altered clock frequencies and/or voltages, will be fit for any particular purpose. Check with memory manufacturer for warranty and additional details.

<Use only if applicable> 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.

<Use only if applicable> 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.

<Use only if applicable> Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

<Use only if applicable> 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.

<Use only if applicable> Intel® AMT should be used by a knowledgeable IT administrator and requires enabled systems, software, activation, and connection to a corporate network. Intel AMT functionality on mobile systems may be limited in some situations. Your results will depend on your specific implementation. Learn more by visiting Intel® Active Management Technology.

<Use only if applicable> Intel is a sponsor and member of the Benchmark XPRT Development Community, and was the major developer of the XPRT family of benchmarks. Principled Technologies is the publisher of the XPRT family of benchmarks. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases.

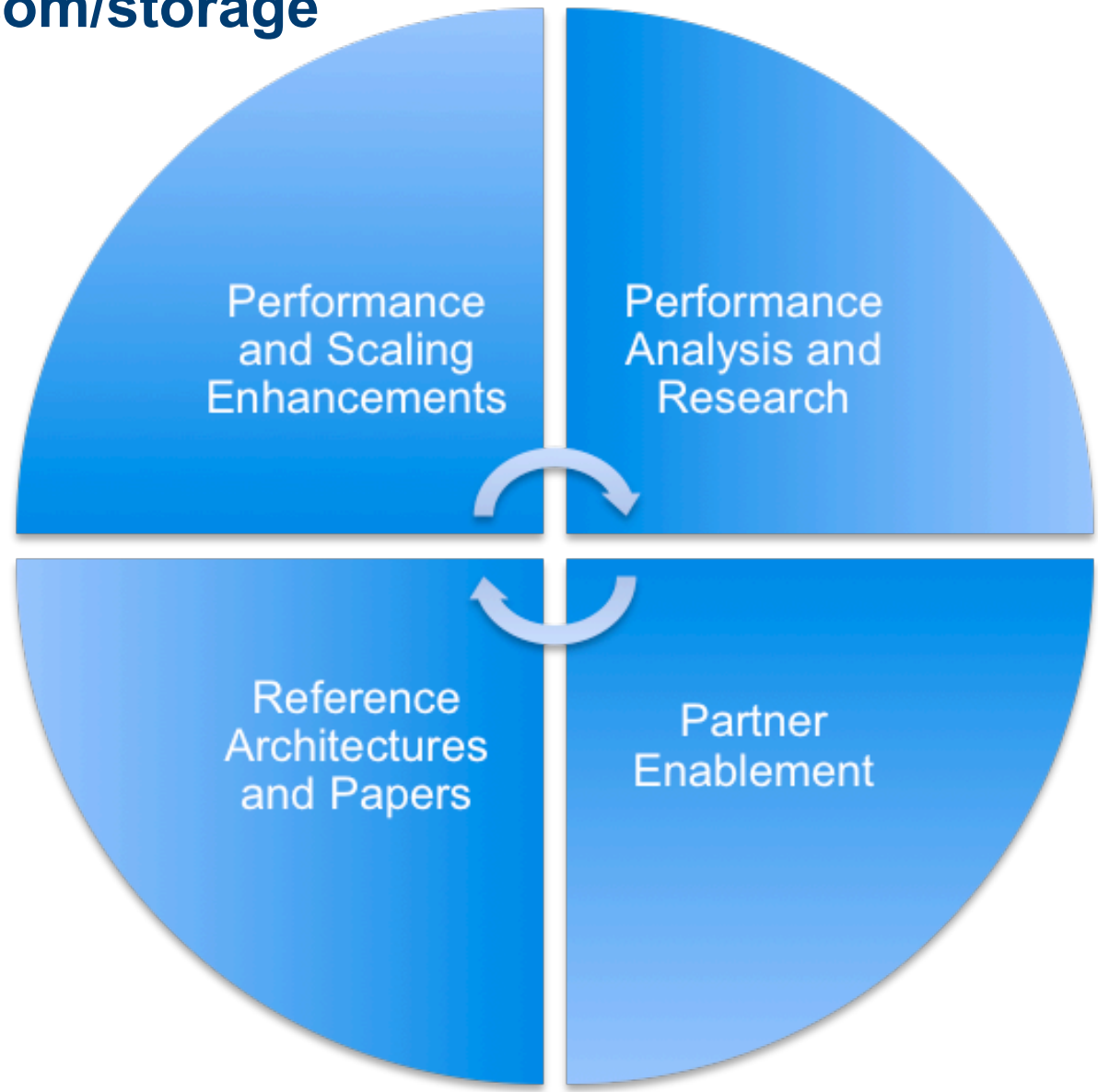
Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

Intel and the Intel logo <Add terms trademarked by Intel and used in this document> are trademarks of Intel Corporation in the U. S. and/or other countries.

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

# Ceph - Areas of Intel Contribution

## see intel.com/storage



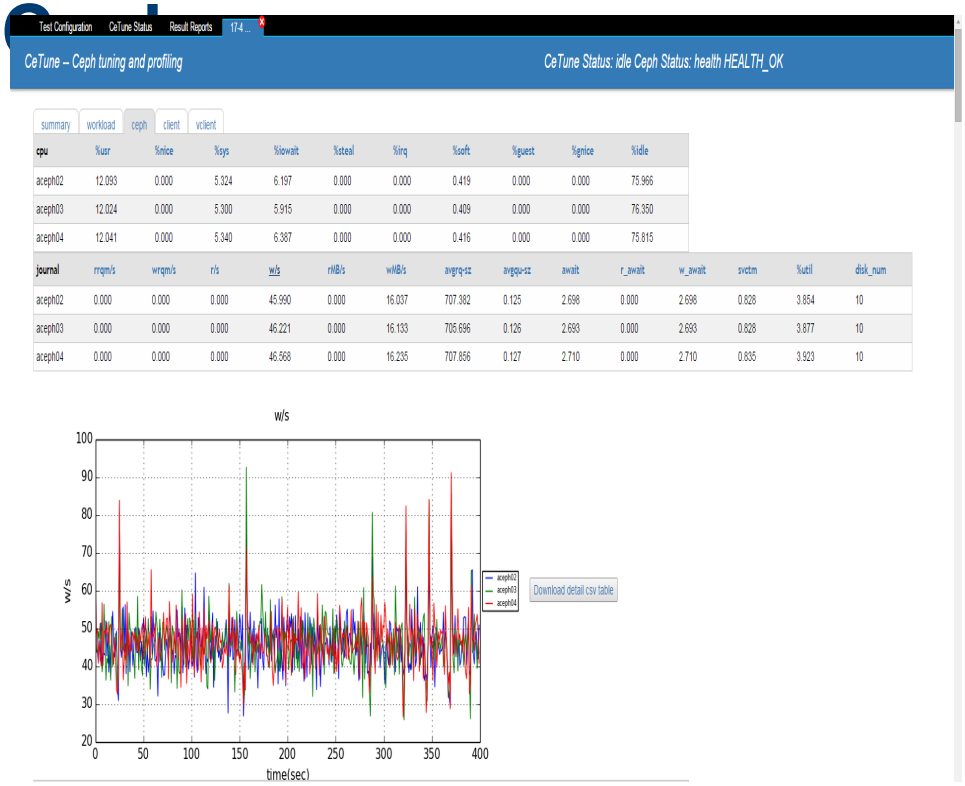
# Ceph Upstream Development Examples from Intel

- Cache tiering proxy read and proxy write
  - Proxy-read: V0.94 HAMMER, Proxy-write: V9.1.0 INFERNALIS RC
- Erasure coding with ISA-L
  - in FIREFLY
- Contributing to Newstore

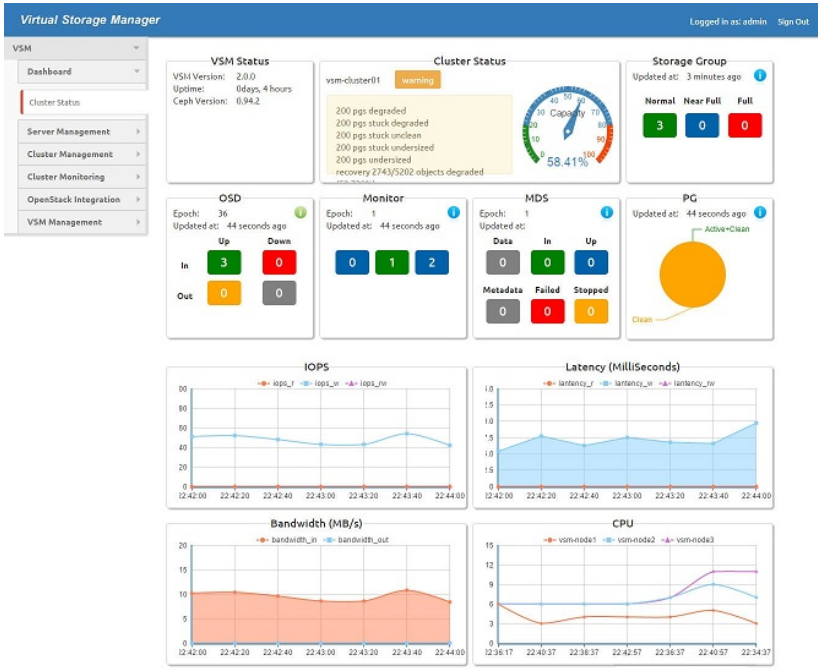
```
[prime][2015-10-26_15-30-25][./count.sh] Caculating Organization commits
 1    4870 Red Hat <contact@redhat.com>
 2     454 Intel <contact@intel.com>
 3     269 XSky <contact@xsky.com>
 4     240 Deutsche Telekom <contact@telekom.de>
 5     202 SUSE <contact@suse.com>
 6     196 Cloudwatt <libre.licensing@cloudwatt.com>
 7     151 Mirantis <contact@mirantis.com>
 8     149 Unaffiliated <no@organization.net>
 9     134 Inktank <contact@inktank.com>
10      68 Reliance Jio Infocomm Ltd. <contact@ril.com>
```

# Open Sourced Ceph Tools (github)

## CeTune (compliments CBT)




## Virtual Storage Manager Manages+Monitors



# Ceph Community Performance Contributions

- Intel and Red Hat hosted Ceph Hackathon with focus on performance optimization, in August
- Intel donated 8 node Ceph community performance cluster named '**Incerta**'
  - One common baseline for performance regression tests and trend analysis
  - Accessible to community contributors
  - Periodic automated performance regression tests with latest builds

| | Sepia Community Laboratory | | | |
|---|---|---|---|---|
| **Cluster Name** | **Incerta** | Plana | Burnupi | Mira |
| **Node Count** | 8 | 95 | 64 | 124 |
| **Node Type** | Intel 2U | Dell R410 | Dell R515 | Supermicro 2U |
| **CPU** | Intel Xeon E5-2650v3 10-Core (2.3GHz) x 2 | Intel Xeon E5620 4-Core (2.4GHz) | AMD Opteron 4184 6-Core (2.8GHz) | Intel Xeon X3440 4-Core (2.53GHz) |
| **Memory** | 64GB | 8GB | 16GB | 16GB |
| **Disk HBA** | On-board | On-board | PERC H700 | Areca ARC-1680 |
| **Disks** | 10 x 2.5" 1TB 7200RPM Seagate ES.2 2.5" SATA | 4 x 500GB WD4 7200 RPM SATA | 8 x 1GB Toshiba 7200 RPM SAS | Up to 8 Various 7200 RPM SATA |
| **NVMe / SSDs** | 4 x Intel 800GB P3700 PCIe 2.5" NVMe | N/A | 1 x Samsung 100GB 2.5" SATA SSD | N/A |
| **Network** | Intel XL710 40GbE QSFP+ | Intel 82599ES 10GbE | Intel 82599ES 10GbE | Intel 82574L 1GbE |
| **Operating System** | CentOS 7.1 | Mixed | Mixed | Mixed |
| **Approximate Age** | Brand New! (Aug 2015) | 3.5 Years | 3.5 Years | 5.5 Years |

*From Mark Nelson @ RedHat: http://permalink.gmane.org/gmane.comp.file-systems.ceph.devel/26635*

- High performance hardware - *3rd Generation Intel Xeon™ E5 Processors, 3.2TB NVMe, 40GbE Networking*
- Supports All HDD, Hybrid (HDD+PCIe SSD), or All PCIe SSD configs

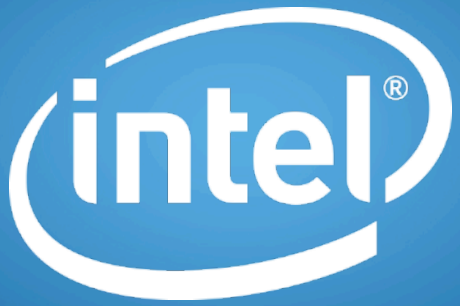
# 4K Random Read & Write on Intel PCIe SSD

| Workload Pattern | Max IOPS |
|---|---|
| 4K 100% Random Reads (2TB Dataset) | **1.35Million** |
| 4K 100% Random Reads (4.8TB Dataset) | **1.15Million** |
| 4K 100% Random Writes (4.8TB Dataset) | **200K** |
| 4K 70%/30% Read/Write OLTP Mix (4.8TB Dataset) | **452K** |

- OSD System Config: Intel Xeon E5-2699 v3 2x@ 2.30 GHz, 72 cores w/ HT, 96GB, Cache 46080KB, 128GB DDR4
  - Each system with 4x P3700 800GB NVMe, partitioned into 4 OSD's each, 16 OSD's total per node
- FIO Client Systems: Intel Xeon E5-2699 v3 2x@ 2.30 GHz, 72 cores w/ HT, 96GB, Cache 46080KB, 128GB DDR4
- Ceph v0.94.3 Hammer Release, CentOS 7.1, 3.10-229 Kernel, Linked with JEMalloc 3.6
  - CBT used for testing and data acquisition
- Single 10GbE network for client & replication data transfer, Replication factor 2



See complete study at http://www.slideshare.net/inktank_ceph


intel