Are Wimpy Cores Good for Brawny Storage?

Calxeda's EnergyCore processors and fabric architecture enable new optimizations for distributed storage appliances.

Executive Summary

When could "wimpy" cores beat brawny ones? They have potential to do so in large-scale distributed storage deployments. This paper explores performance and resiliency tradeoffs enabled by using Calxeda's ARM-based EnergyCore processors and their fabric-based system level architecture as the underpinning for a Ceph distributed object store implementation.

Calxeda, Inktank, and Penguin announced a partnership in June, 2013 to deliver a software-defined storage appliance based on Calxeda's EnergyCore processors and hardware architecture, Inktank's Ceph distributed storage software architecture, all on Penguin's UDX1 servers.

Based on performance comparisons run by Calxeda, Ceph scales well using Calxeda's processors and system architecture. Calxeda is currently working with customers and partners to create a performance measurement framework for comparison to x86 storage servers.

Distributed Storage Background

Big Data and cloud-computing are creating demand for large-scale, resilient storage systems. These storage systems are distributed and are designed to avoid the pitfalls of traditional storage systems. Distributed storage systems seek to overcome scalability problems associated with table-based index lookups, single points of failure, and complex hardware resiliency schemes such as RAID (redundant array of inexpensive disks). File systems are a subset of these more general storage architectures: distributed storage systems typically offer block, object, and file-based storage interfaces.

Distributed storage architectures are built from "storage servers." Storage servers are throughput-based server appliances that host some amount of local DAS (direct attached storage). Like NAS (network attached storage) and SAN (storage area networks) in traditional enterprise IT environments, distributed storage systems run their own OS and dedicated workload. Thus storage servers do not have to run the same OS as the rest of an at-scale server farm; they can even run on entirely different processor architectures from the compute systems.

However, there are massive differences in scale between enterprise IT's "finite" storage requirements and hyperscale services' "at-scale" storage requirements. SAN systems require that each server in a group of networked servers manage their own file systems on a network-remote virtual file system. While NAS systems implement a dedicated server appliance and can run their own local applications, they present a single interface to the network, which doesn't scale well. Hyperscale datacenter customers want to run storage apps, such as replication and de-duplication, closer to the physical disks and absolutely require the aggregate storage bandwidth to scale as more storage is added.

The benefits of running storage apps closer to the physical disks are reduced storage network traffic and power, and increased storage system responsiveness and resiliency.

Distributed Storage Systems

A sampling of hyperscale distributed storage systems includes:

- **Inktank's** Ceph
- **OpenStack's** Swift
- **Red Hat's** Storage Server (formerly GlusterFS, acquired October 2011)
- **EMC's** ScaleIO ECS (acquired July 2013)

HP's 3PAR solutions and **Dell's** DX Object Storage both have partnerships with a large cross section of the "open" distributed storage world, including Inktank, and both have invested substantially in OpenStack and Swift.

Inktank's Ceph is compelling. Not only is Ceph open source, it has been integrated with OpenStack and CloudStack operating environments. OpenStack is complimentary toward Ceph as a replacement for Swift, saying:

> *"Ceph's advantages are that it gives the administrator more fine-grained control over data distribution and replication strategies, enables you to consolidate your object and block storage, enables very fast provisioning of boot-from-volume instances using thin provisioning, and supports a distributed file system interface."* [Source here, with caveat – the distributed file system interface is not yet ready for production deployment].

Ceph Background

The core of Ceph is RADOS (Reliable, Autonomous, Distributed Object Store). RADOS uses an algorithm Inktank calls CRUSH (Controlled Replication Under Scalable Hashing) to compute data locations on-demand, instead of keeping a central lookup table. Lookup tables don't scale very well at cloud scale. Ceph doesn't have to remember where it stores data, because it can easily calculate that information. Because storage locations are calculated, Ceph can implement intelligent data replication, can transparently move data closer to where it is being consumed, and can

heal itself when storage nodes are disabled. Ceph accomplishes all this automatically and at cloud scale.

RADOS is built from three key components: Object Storage Device (OSD), Monitor, and Client. (Because Monitors are not directly involved in the data flow between OSDs and Clients, we'll defer discussion of them. In other words, Monitors are not performance sensitive and run as well on ARM-based systems as x86-based systems.)

A Ceph OSD is an independent executable that runs on a storage server processor. The OSD manages its own local file system and a logical disk used to store its files. It stores all of its data as objects in a flat namespace, which means there are no folders or directories.

Ceph recommends one OSD per physical drive. A hard disk drive (HDD) or solid-state drive (SSD) is preferable, but RAID setups are also possible. Ceph supports from tens to hundreds of thousands of OSDs in a storage cluster. OSDs also peer with each other to perform the intelligent, higher level tasks (such as replication and recovery, which are initiated by the Monitors mentioned earlier) that enable Ceph to operate at cloud scale.

Ceph is designed to store at least two copies of each object with each copy on separate physical media for file recovery when physical media fails. *When* physical media fails, not *if*: the statistical probability scales with the size of the storage cluster, and in large clusters the probability is high enough that it is an expected occurrence. When a storage server node fails, all of the files stored in all of the drives attached to it will then be unreachable until the server node is repaired. Ceph tracks which OSDs are running on the same processor (or pool of drives) and adjusts its replication strategy so that redundant files are not stored on disks associated with the same processor socket.

One of Ceph's really interesting design philosophies is that it pushes the compute component of storage

PENGUIN ON DEMAND (POD)

POD is a public, pay-as-you-go HPC cloud. Penguin recently added their Calxeda EnergyCore-based UDX1 storage server system to POD to support their Hadoop and HPC products with a Ceph distributed storage capability.

Each UDX1 contains 12 EnergyCore SoCs, for a total of 48 ARM cores. The chassis hosts up to 36 HDDs or SSDs – 3 drives per SoC, giving each drive a dedicated ARM core to run a Ceph OSD with spare cores in the design for resilience. Ceph now accounts for about 20% of POD's storage mix.

David Ingersoll, Penguin Worldwide VP of Sales, says that they are seeing 2x power efficiency gains by implementing Calxeda-based storage servers, and that Penguin's customers are showing tremendous interest in evaluating those servers via POD – 32-bit ARM cores satisfy this storage use case today.

closer to the storage media to prevent single points of failure. OSDs' peer-to-peer nature allows a small set of fairly simple rules to express robust and complex features – an apt example of emergent behavior.

Calxeda EnergyCore and Ceph

EnergyCore SoCs have several advantages in a Ceph storage system. Low power and small motherboard area requirements enable system designs that increase the ratio of SoCs per drive without increasing their power budget and design complexity. When that happens, the first order effects are more I/O performance per disk and the ability to dedicate a processor core to each OSD/drive pair.

In addition, the EnergyCore fabric's Ethernet switching capabilities enable cost effective east-west, peer-to-peer bandwidth for storage architectures, further increasing available system bandwidth and improving storage throughput.

One of the most important benefits of using EnergyCores in a Ceph storage cluster is the inherent small size of individual failure domains. If an EnergyCore-based storage server node were to fail, a limited number of disks would be affected. Thus a smaller number of file/object replications would take place to restore the correct number of copies lost for each object on each drive attached to the failed node.

Calxeda's storage partnerships and design-wins at Foxconn and Aaeon are aimed directly at this distributed storage market.

Ceph Performance Using Calxeda EnergyCore Processors

Calxeda recently conducted a small set of Ceph I/O performance measurements*. They varied the number of processor nodes in a system (6, 8, 10, 12, and 36) and the number of hard disk drives (HDD) attached to each node (2 or 3). This type of self-comparison is useful for illuminating how well an architecture scales. It is intended to answer the question: "Do we see linear improvement when we add more resources?"

When we cut the raw data by drive and by node count, we do see linear scaling effects.

The take-away from the following charts is that the aggregate storage bandwidth of these Calxeda-based Ceph system configurations does increase in a predictable fashion as both the number of nodes increases and as the number of HDDs increases.

Figure 1: Total System Throughput by Node Count

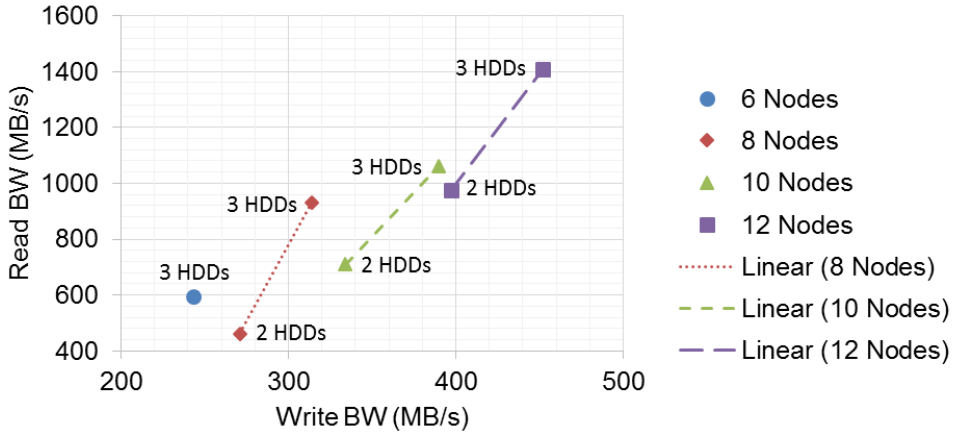
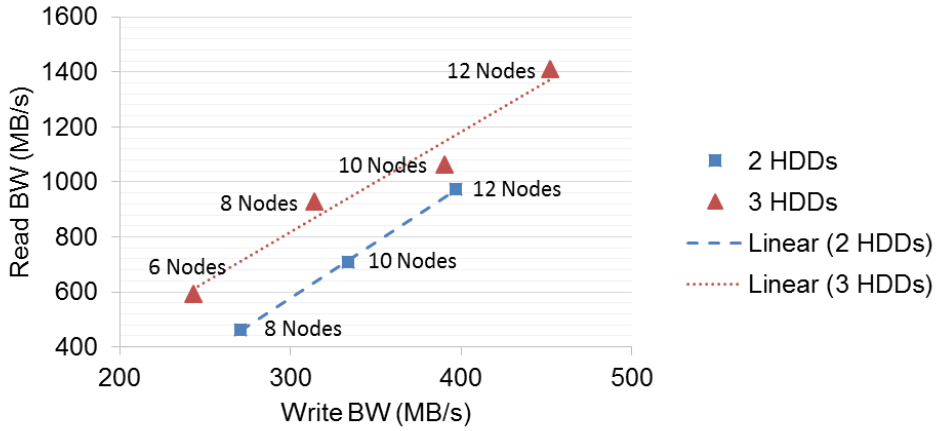


Figure 2: Total System Throughput by Drives per Node



[Data for 36 nodes omitted from above charts for legibility.]

These charts normalize the I/O bandwidth results by the number of nodes in a system, and then by the number of drives in a system:

Figure 3: System Throughput per Node

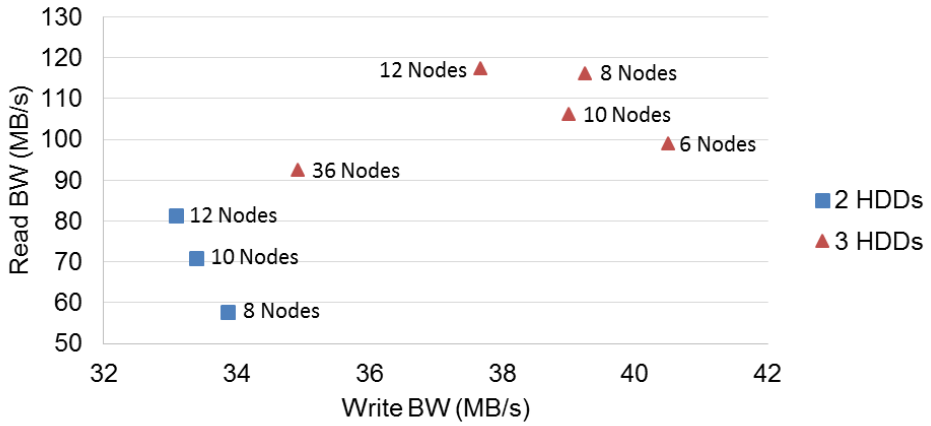
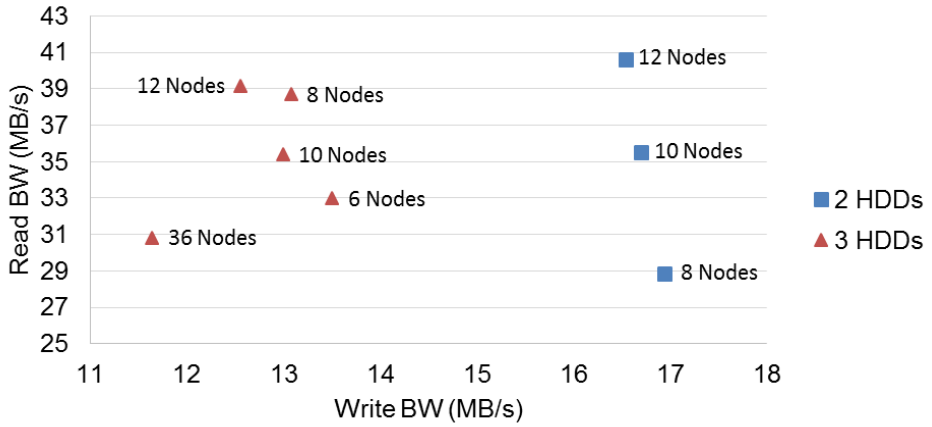


Figure 4: System Throughput per Drive



[The three-drive data point standing somewhat alone is the 36-node system, which begins to show non-linear scaling effects. However, it is still loosely clustered with the rest of the three-drive data points. We removed the 36-node data point from the following general conclusions.]

Moving from two to three drives, of course, drives higher overall system read and write bandwidth. In an ideal system that scales completely linearly when adding drives, we would expect read and write performance both to scale 50% by adding a drive to a node. When we separate throughput by the number of drives in a system, we see on average a 57% increase in read bandwidth and a 17% increase in write bandwidth. Flash caches on each drive might enable higher read bandwidth scaling, and those effects can be addressed in future performance analysis. Write bandwidth scales below the ideal, because adding another disk with equally slow write performance to a node is more than likely saturating each node's data cache with diminishing returns.

The story is slightly different when system performance is normalized to a per processor node performance story. In that scenario, it is clear that read bandwidth scales with node count, but write bandwidth per node is static within each of the disk counts. There is, on average, a 28% write performance improvement at the node level by adding a third disk. We believe that write bandwidth is predominantly affected by system data caching and disk write performance.

Future analysis should examine the effects of adding more memory at each node, using differently configured HDDs, using solid state drives (SSD) instead of HDDs, and of course running different workload profiles to drive different throughput characteristics.

Call to Action

Calxeda's EnergyCore architecture enables Ceph to scale well and address a range of throughput performance, while also demonstrating at-scale failure resiliency and tighter binding of compute and storage resources for the Ceph distributed object store.

Sage Weil, creator of the Ceph project, says that from a development perspective Ceph's developers cannot tell that their Calxeda-based storage servers are any different from x86-based equivalents; – there is no difference in tools, build process, etc. From a performance point of view, EnergyCore SoCs enable Ceph to credibly scale their distributed storage network.

If you are considering deploying a highly scalable distributed storage system, consider running Inktank's Ceph on Calxeda EnergyCore based hardware. Calxeda and Inktank are working together to optimize Ceph on Calxeda-based hardware, and that collaboration is creating a significant new capability for hyperscale datacenter deployments.

Resources

Calxeda Big Data
Inktank What is Ceph?
Calxeda and Inktank Press Release
Penguin, Calxeda and Inktank Press Release
Inktank Ceph on ARM Demo Video
Penguin on Demand (POD)
MIS Big Data is Extra Sensory Correlation

* System Configuration Used for Performance Analysis

Node Definition

- A "node" is a single system-on-chip (SoC) plus memory
- Single (1) chassis used for 6 to 12 nodes
- Three (3) chassis used for 36 nodes, each chassis containing 12 nodes

Calxeda Systems

- One (1) to three (3) Penguin UDX-1 chassis
- Three (3) EnergyCard RB101A004-4414 processor cards per UDX-1 chassis
- Four (4) ECX-1400 SoC (1.4Ghz) per EnergyCard
- Four (4) 1GB DDR3L-1333 ECC DIMM cards per EnergyCard
- Two (2) to three (3) Seagate ST3000NC002-1DY166 HDDs per EnergyCard (3TB, SATAv3, 3.5", 7200rpm)

Network and Load Testing Harness

- Unreleased 48-port whitebox 10Gbps switch
- UDX-1 chassis connected via two (2) 10Gbps Ethernet links (LACP-enabled)
- Five (5) whitebox load testing "client" servers connected to switch via one (1) 10Gbps Ethernet link each

Inktank Software

- Ceph "Cuttlefish" v0.61.1 released May 9, 2013
- Ceph "radosbench" tool running on client servers to generate load/demand

Author
Paul Teich, Senior Analyst at Moor Insights & Strategy

Editor
Patrick Moorhead, President & Principal Analyst at Moor Insights & Strategy

Inquiries
Please contact us here if you would like to discuss this report and Moor Insights & Strategy will promptly respond.

Citations
This note or paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

Disclosures
Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies, including Calxeda, who commissioned this paper. No employees at the firm hold any equity positions with any companies cited in this documented.

Disclaimer
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness or adequacy of such information and shall have no liability for errors, omissions or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.