Building All-Flash Software Defined Storages for Datacenters

Ji Hyuck Yun (dr.jhyun@sk.com)

Storage Tech. Lab

SK Telecom

Synergy between SK Telecom and SK Hynix

SK telecom

SK broadband

SK planet

Service & Solution Provider

Mobile & Fixed Broadband Service
#1 Carrier in South Korea

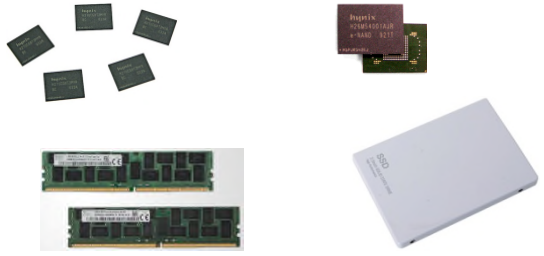
cloud tv map

cloud biz ST syrup

SK Hynix merged into SK Telecom at Feb. 2012

SK hynix

Memory Solution Manufacturer



Develop, Operation, and Management

National Network Infrastructure

Datacenter IT Infrastructure

Reducing TCO, improving delivered QoS

R&D on ICT Infra & System

Server, Network, Storage, Platform, S/W-defined x ...

Real-world workloads & benchmarks

Cloud Computing, Big Data, and Video Technologies selected to leverage the effectiveness of flash

【 Major Trends 】

Cloud

Flexible, Scalable, Agile Infra
: Scale-out Storage

Big Data

Massive Information
: Data Analytics

Video

High Quality Contents
: Video Streaming

SK telecom SK broadband SK planet

Cloud, Video, Big Data Infra
With Real workloads

Workloads, Unmet needs, Pilot Test

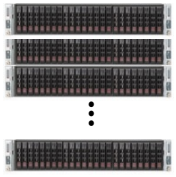
"Vertically Optimized System Enhanced by Flash Tech"

① All-flash Scale-out Storage for Cloud Infra	② All-flash CDN Server for Video Streaming	③ Big Data Analytics System Accelerator

S/W-defined Platform	CEPH	CDN	Spark Hadoop
System/Infra HW Architecture			
OS Kernel SW			
Device (RTL/Firmware)			
Memory Semiconductor			

All-flash Scale-out Storage

In the enterprise, both scale-up & scale-out storage are evolving to all-flash to meet with increasing performance requirements of recent applications

	HDD or Hybrid	All-flash
Scale-up Storage	▪ EMC, NetApp, etc 	▪ Pure Storage, etc  • Traditional Infra Target
Scale-out Storage	▪ CEPH*, etc  Commodity Server H/W + Scale-out Storage S/W (Minimizing Infra TCO)	▪ SolidFire, etc  ▪ Cloud Infra Target ➢ SKT is optimizing CEPH by customizing storage and network S/W stack for flash Commodity Server H/W w/ All SSD (SKH Commodity 2.5")

All-flash Evolution for Performance Maximization

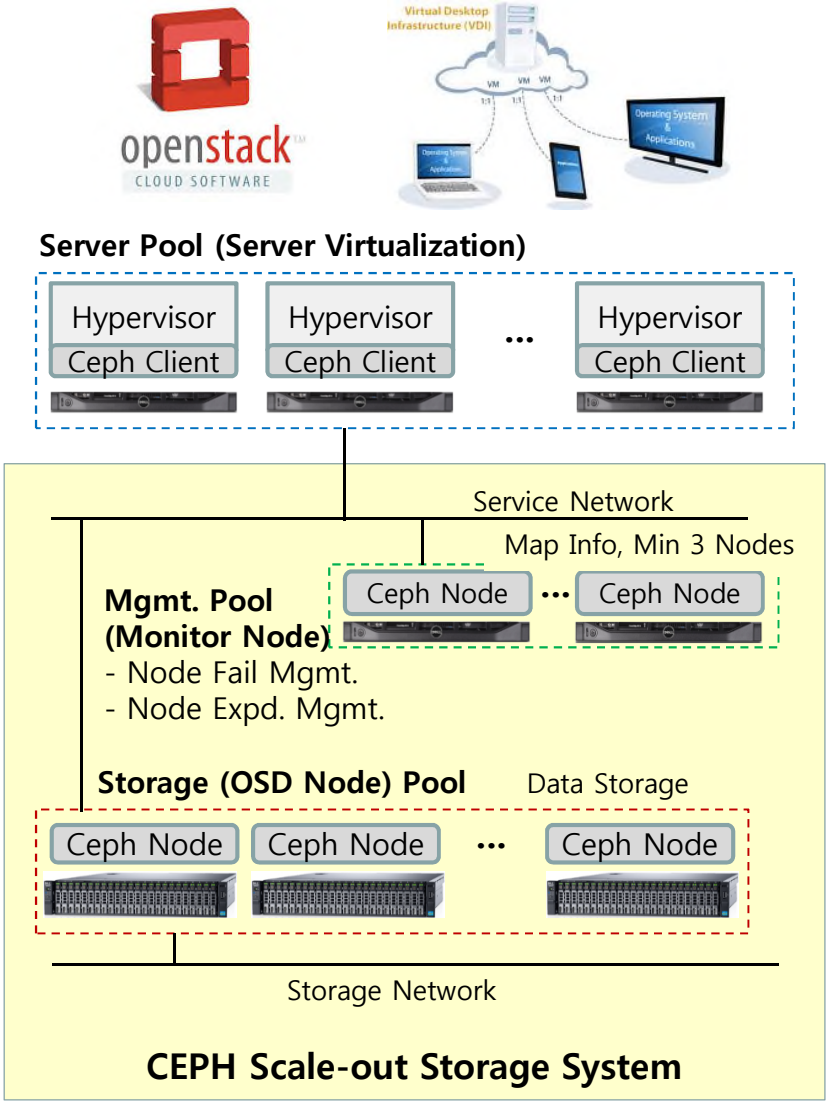
Low Cost Scale-out Architecture for Cloud Infra

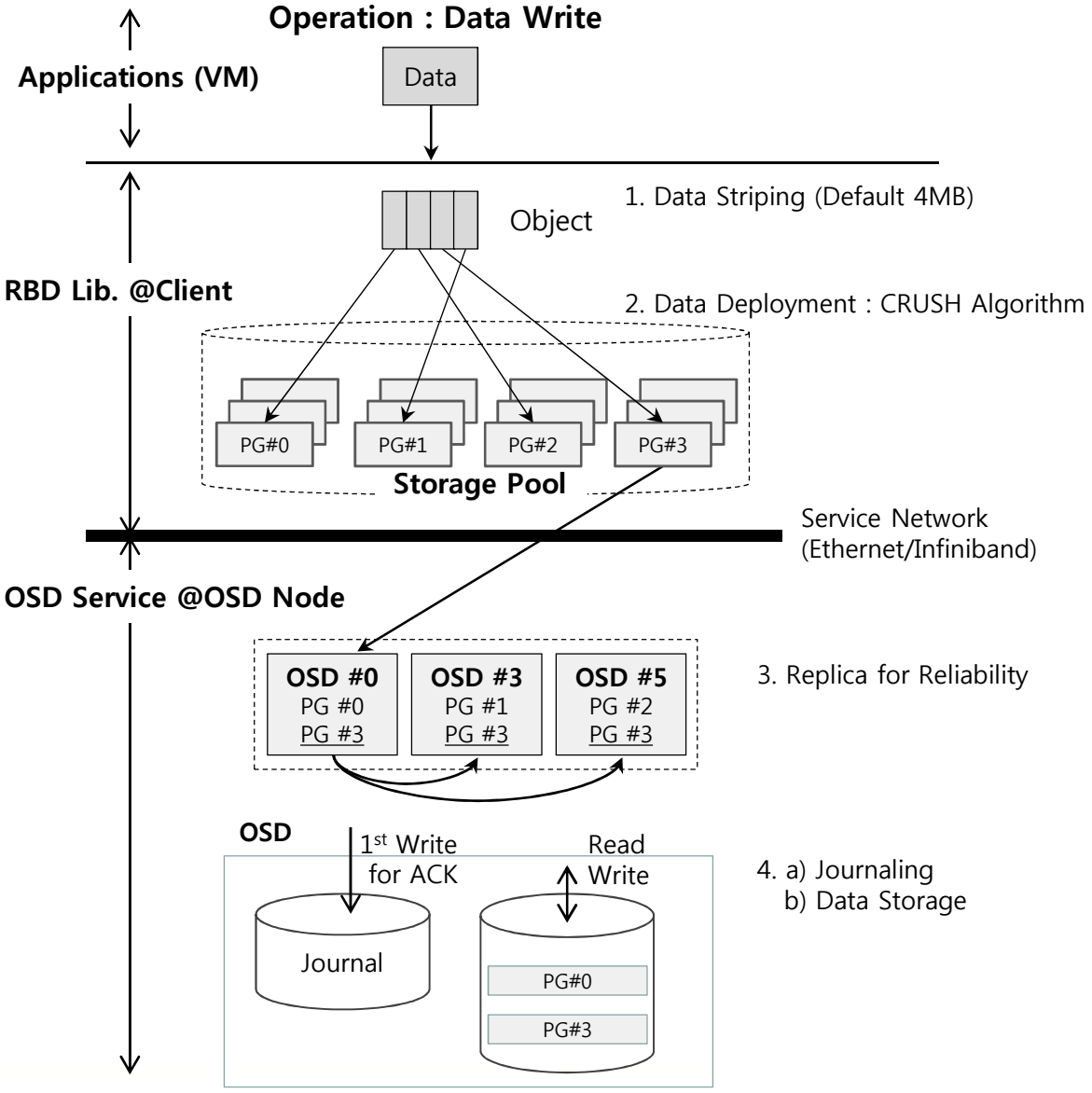
Targeting to SKT SDDC Infra
(Openstack based Private Cloud)

➢ **Infra Cost ↓**
➢ **Scalability ↑**

*CEPH: Open-source scale-out storage S/W to present object, block, and file storage from distributed servers

CEPH provides scale-out capacity expansion but performance is inherently restricted because of its HDD centric design





SK Telecom CEPH system consists of 1U servers for a small footprint, and employs an NVRAM–SSD hierarchy to achieve high performance.

Configurations : 4 Node + Microserver 1 set

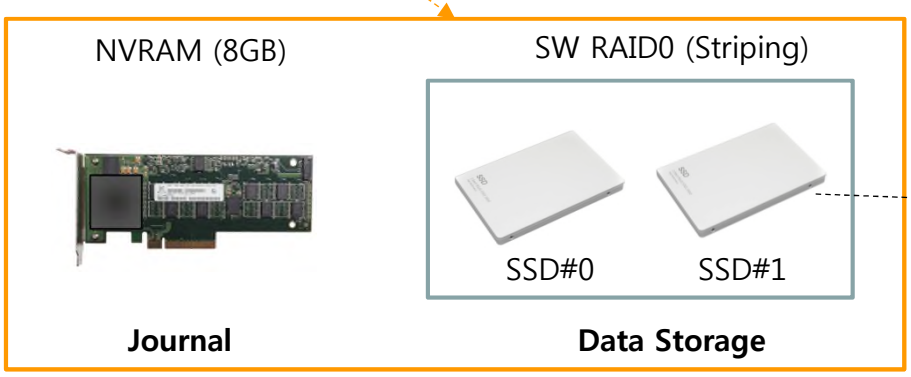
OSD Node

Monitor Node (3ea)

2U

SKT Microserver
- CPU: Intel Rangeley C2738 (2.4GHz/8C)
- DRAM: 32GB/1600MHz
- Network: 10GbE/1GbE

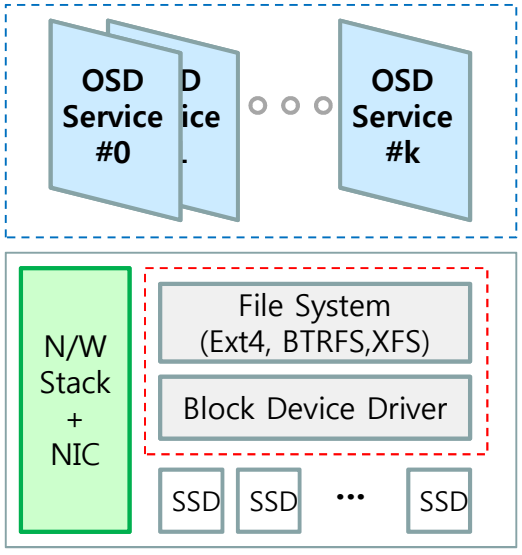
`15.12 Target CEPH Node H/W Spec.
- CPU: 2x E5-2690v3
- DRAM: 128GB /2133 MHz
- Network: 10GbE for Service, 10 or 40GbE for Storage
- Replication: 2
- Storage: NVDIMM (8GB) for Journal,
  SSD for File Store



OSD Configuration

SKH SATA SSD
- 1TB (or 2TB) capacity based
- Physical Capacity per Node: 10TB

SK Telecom has done an in-depth analysis of CEPH, and has implemented customizations and optimizations to better exploit the performance of flash



Optimizations

1. **Light-weight Ceph Service**
2. **Efficient Object Storage**

`15

OSD Optimizations
- Lock, thread, priority ....

NVRAM/NVDIMM Deployment

High Performance Network & Protocol
- 40GbE, infiniband, RDMA, etc.

`16~

Non-File System Key/Value Storage
(Key/Value APIs, Proprietary SSD)

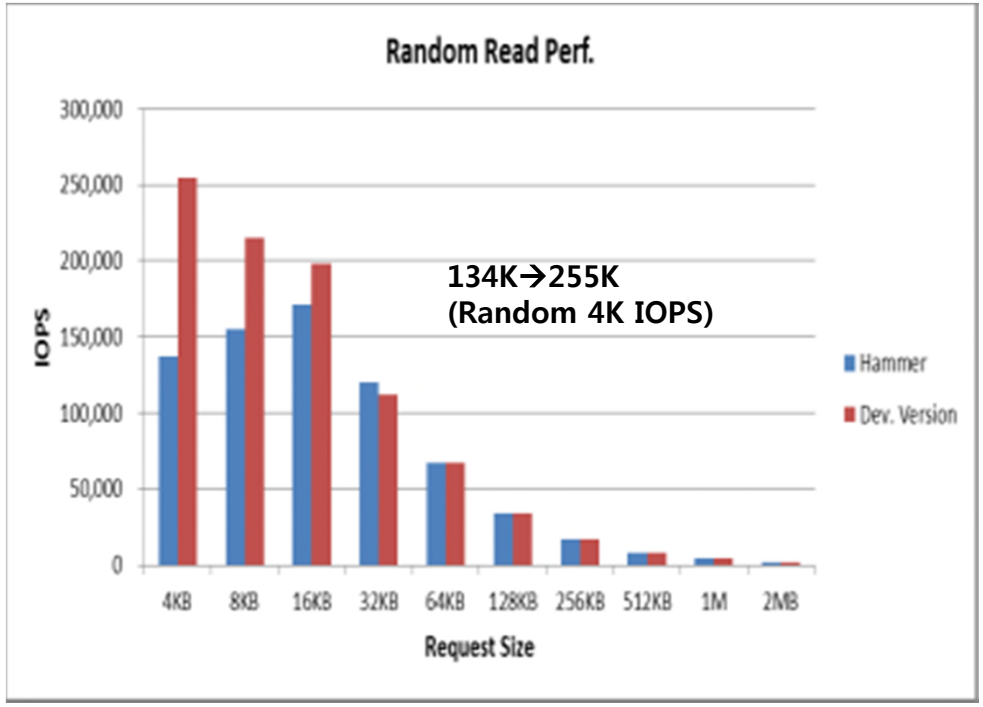
- File-system Overhead Elimination
- Efficient meta-data mgmt.

Introducing NVMe Flash SSD

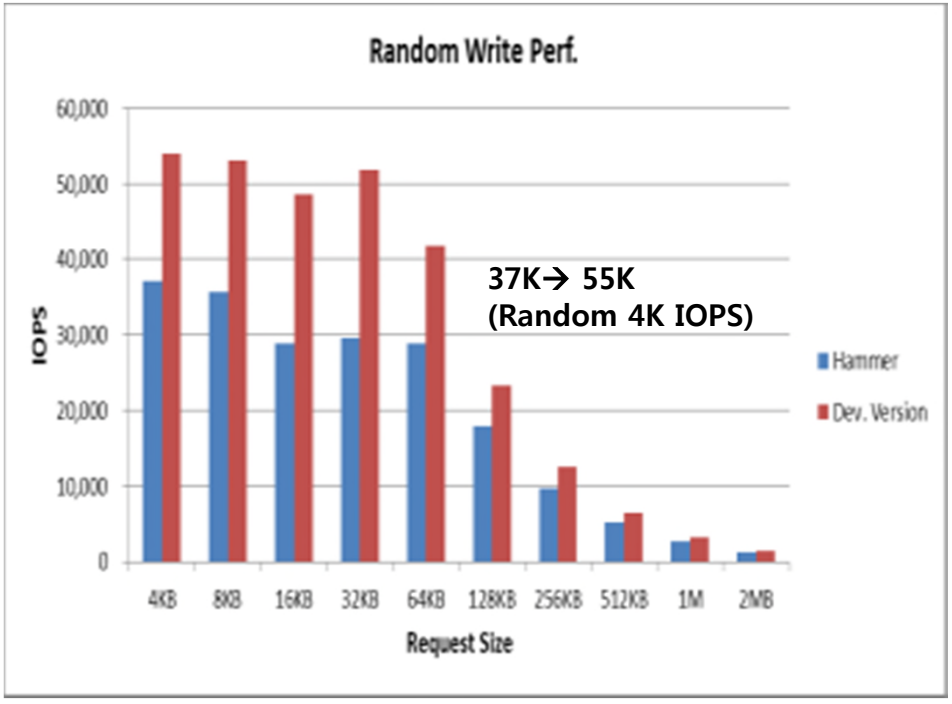
All-flash Scale-out Storage

Evaluation Results

<Read Performance>



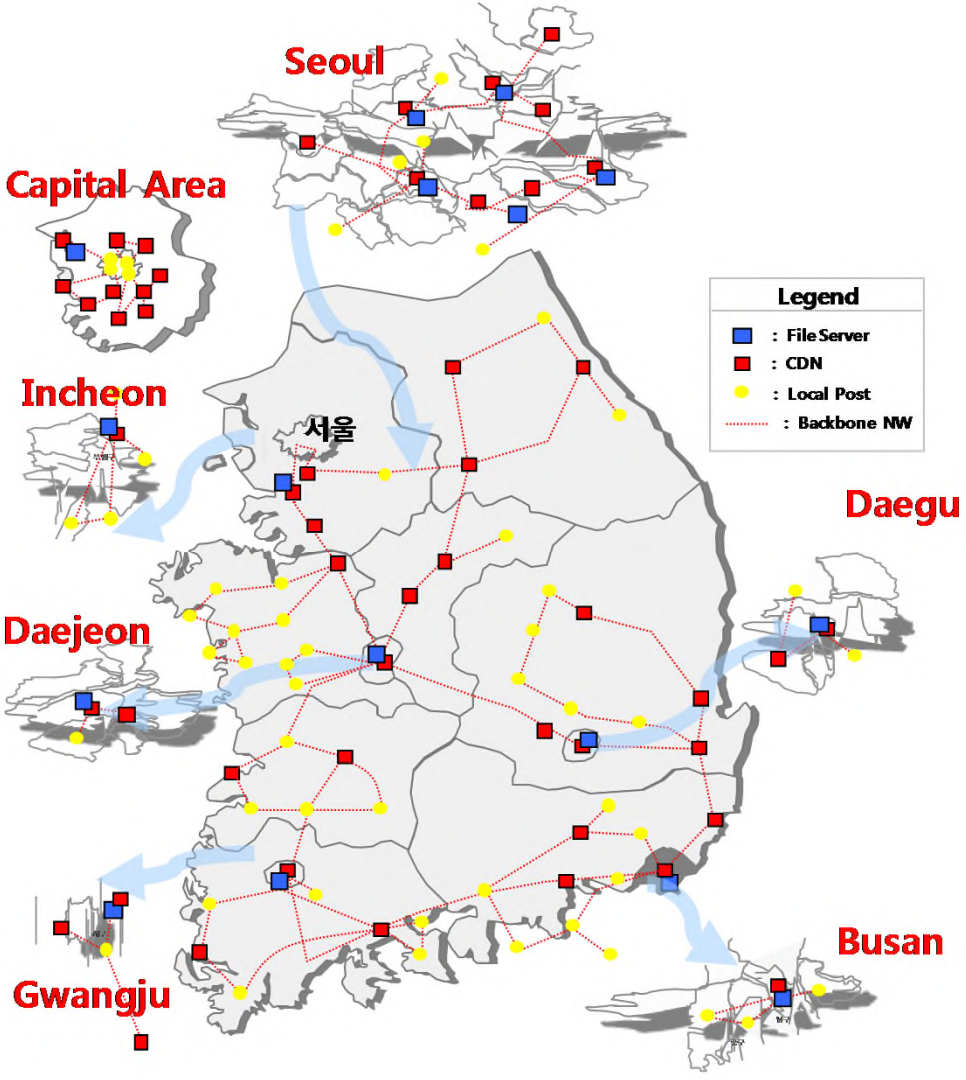
<Write Performance>



We are open to collaboration with the goal of achieving maximum performance in all-flash CEPH systems

All-flash VoD Streaming System

【 Example of Content Delivery Network Infra 】



- Hot contents are cached geographically at local sites to reduce overall network traffic and to guarantee on QoS of VoD Streaming.
- To cache video contents, each local site has CDN server clusters. The number of CDN servers is determined in order to meet the maximum required service bandwidth of customers living in that area.
- As the number of customers increases and video contents quality become better, the number of CDN servers and capacity of CDN servers grow simultaneously.

CDN Server R&D Target

Infra TCO Reduction can be achieved by

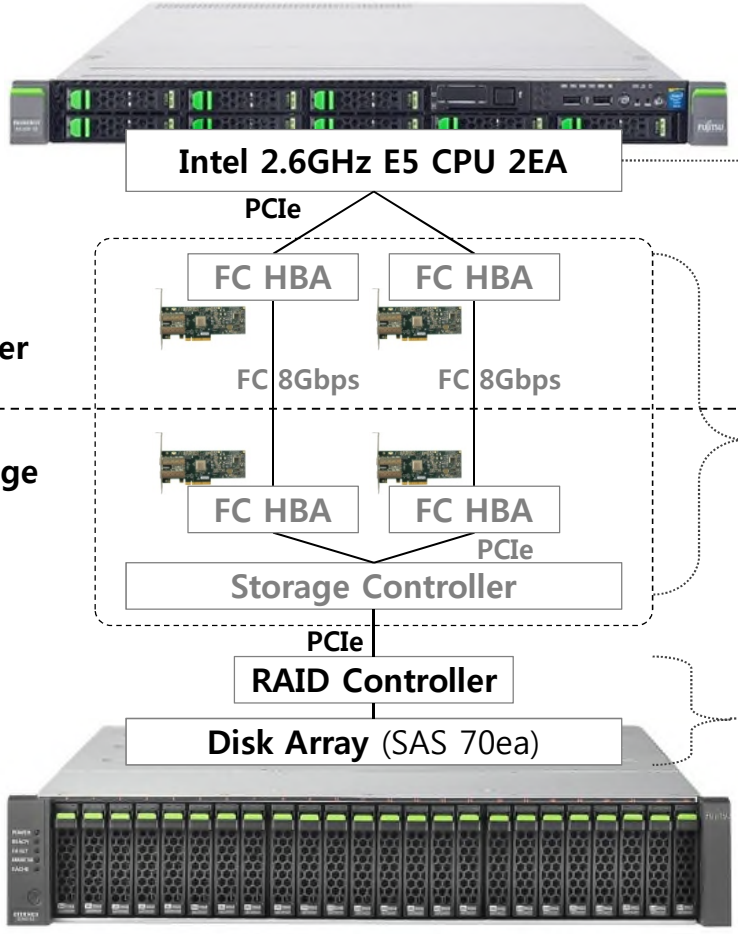
- Service Bandwidth Increase
- Server BOM Reduction
- Power consumption reduction
- Footprint reduction

All-flash VoD Streaming System

System Architecture Optimization

High performance, high density, low power, and low cost are achieved by eliminating redundant components in I/O path, PCIe connected flash, and utilizing an optimal CPU.

【 Legacy CDN Server Architecture 】



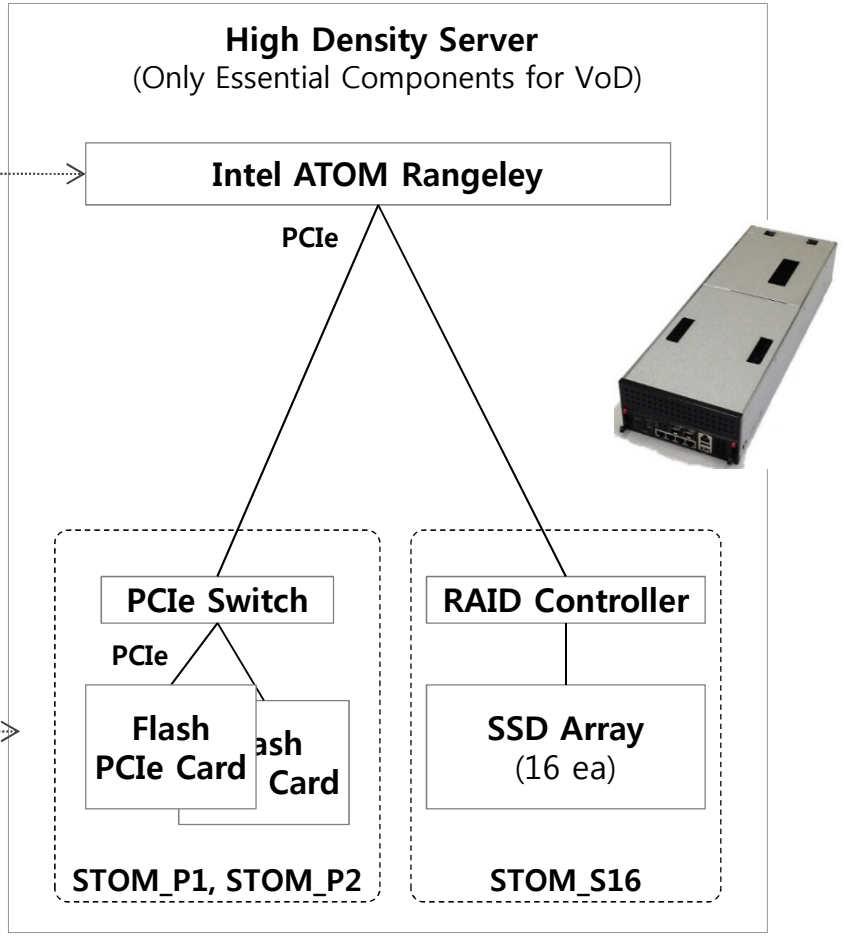
In-service CDN (Edge) Server Architecture (3.3Gbps Service)

Optimization

Elimination

Optimization

【 SKT STOM CDN Server Architecture 】



All-flash VoD Streaming System

SKT STOM series consists of P1(5TB), P2(10TB), S16(14TB) models, supporting various requirements: high performance/high density/low power/low cost.

- ❖ **High performance of all-flash**
  - ✓ **Fast Retrieval of Video Contents**
  - ✓ **Capacity: 5TB~16TB**
- ❖ **Dedicated appliance for VoD**
  - ✓ **Component optimization for VoD**
- ❖ **Flexibility of density/capacity/cost**
  - ✓ **NVMe PCIe Flash Card 1EA/2EA, 2.5" SATA SSD 16EA**

Model	Form factor	Storage Support	Main Features (For Each Node)
STOM_P1 (Performance| Density)	**4 Server Nodes in 2U** (0.5U Server effectively) 	**Flash PCIe Card 1EA / Each Node**	**Storage Capacity:** - 5TB (w/ 1x 5TB PCIe Card) **Network Interface:** - 4 x 1Gbe Copper ports - 2 x 10Gbe Fiber ports **Max. Bandwidth:** - HTTP: >18Gbps
STOM_P2 (Performance| Capacity)	**2 Server Nodes in 2U** (1U Server effectively) 	**Flash PCIe Card 2EA / Each Node**	**Storage Capacity:** - 10TB (w/ 2x 5TB PCIe Card) **Network Interface:** - 4 x 1Gbe Copper ports - 2 x 10Gbe Fiber ports **Max. Bandwidth:** - HTTP: >18Gbps
STOM_S16 (Capacity|Cost)	**2 Server Nodes in 2U** (1U Server effectively) 	**2.5" SATA SSD 16EA / Each Node**	**Storage Capacity:** - 8TB (w/ 16x 0.5TB SSD) - 16TB (w/ 16x 1TB SSD) **Network Interface:** - 4 x 1Gbe Copper ports - 2 x 10Gbe Fiber ports **Max. Bandwidth:** - HTTP: >18Gbps

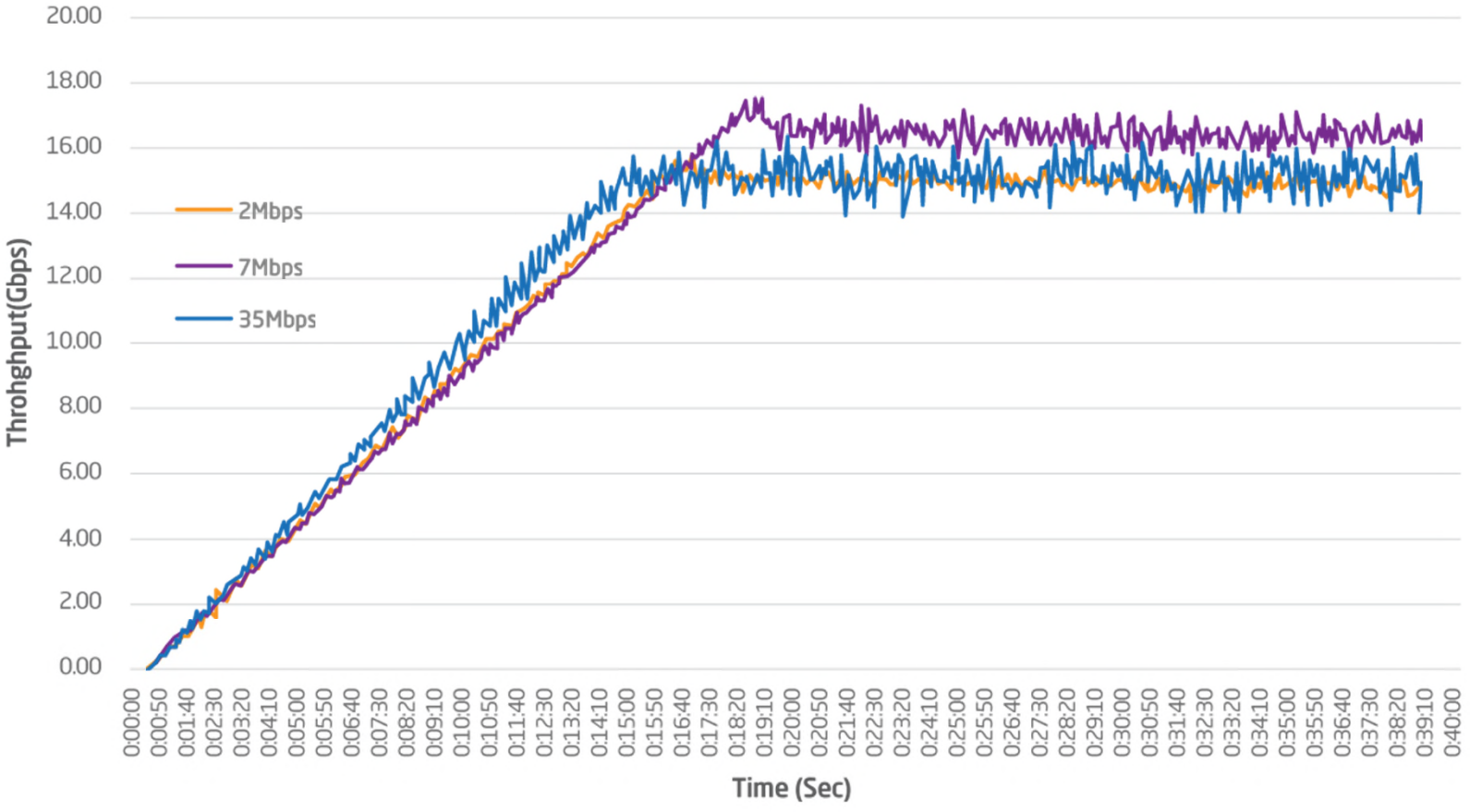
SKT STOM servers can cut power consumption by 90% and cost by 40% compared to a HDD-based CDN server

	HDD CDN Server (`15)	SKT STOM: All-flash CDN Server (`15)			
Footprint (Size)					
		STOM-A_P2		**STOM-A_S16**	
Computing	Intel E5	*Power & BOM reduction*	Intel Atom Rangeley	*Power & BOM reduction*	Intel Atom Rangeley
Max Service BW	10Gbps(RTSP)	~ 10Gbps(RTSP) ~20Gbps(HTTP)		~ 10Gbps(RTSP) ~20Gbps(HTTP)	
Storage Module	300GB HDD 70EA (15TB*)	*Performance focused*	SKT PCIe Card 5TB 2EA (10TB)	*Capacity & Cost focused*	SKH SATA SSD 1TB 16EA (14TB*)
Streaming Contents	Full HD	Full HD Ultra HD		Full HD	
Power Consumption	200 W/Gbps (Estimated)	*90% ↓*	17 W/Gbps	*95% ↓*	11 W/Gbps

* Total Usable Capacity by RAID Configuration

2 All-flash VoD Streaming System

STOM Performance



Hynix SSD	2Mbps	
	User	Throughput
	8,100	14.78
	7Mbps	
	User	Throughput
	2,400	16.43
	35Mbps	
	User	Throughput
	430	15.08

※ Users : max. # of simultaneous users at each bit rate, Throughput: Gbps

- Video clients were emulated using HP LoadRunner performance analysis package
- Throughput was measured for video streams encoded at increasing bit rates: 2Mbps (HD), 7Mbps (FHD) and 35Mbps (UHD)

Next generation of STOM exploits state-of-art Intel CPUs and provides higher density storage

STOM-A Family

Intel Atom Rangeley SoC

STOM-A P1 (`15.2)
Max. BW: - RTSP: >9Gbps - HTTP: >16Gbps
Storage: - 1x PCIe Card

STOM-A P2 (`15.3)
Max. BW: - RTSP: >9Gbps - HTTP: >16Gbps
Storage: - 2x PCIe Card

STOM-A S16 (`15.4)
Max. BW: - RTSP: >9Gbps - HTTP: >16Gbps
Storage: - 16x 2.5" SATA

Intel High Performance Enterprise CPU (`15.3 Release)

- ✓High Performance XEON Architecture
- ✓14nm Tech, DDR4 Memory
- ✓Cost/Density Competitive Edge (Identical Design)

STOM-D Family (`15.4Q)

Intel Xeon D SoC

STOM-D P2
Max. BW: - RTSP: ~20Gbps - HTTP: ~40Gbps
Storage: - 2x PCIe Card

STOM-D S16
Max. BW: - RTSP: ~20Gbps - HTTP: ~40Gbps
Storage: - 16x 2.5" SATA

※ Max. BW: Expected

x2 Application Performance

❖ Ultra High Performance/Massive Capacity CDN Server : STOM-E (`16.1H)

- Max. CPU Performance (Encryption, etc.), Wide N/W (80Gbps), Massive Storage, Space Efficiency (1U)
  → Computing (Xeon-D or Xeon-E5) & Storage (SATA SSD or NVMe SSD) Modular Architecture

3 Big Data Analytics Accelerator

Enhancing T-DW (SK Telecom's Hadoop-based Data Warehouse solution) **by employing SSDs with customized caching SW**

T-Data Warehouse

Ex) Big Data from N/W Equipments are analyzed to operate and mgmt. infrastructure more effectively and make N/W stable to the limit.

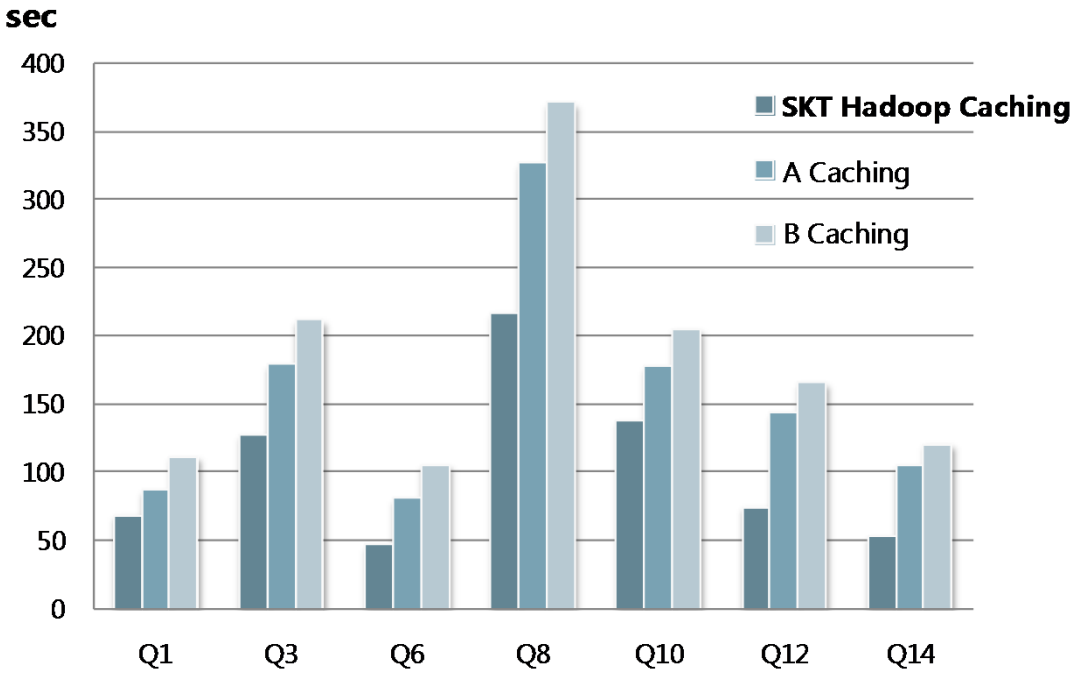
SSDs are applied with customized caching SW to workloads

SSD to each server for cache

Applied to SK Telecom infrastructure in `15

Hadoop/Spark based cost-effective solution

TPC-H Comparison to conventional flash caching SW in Hadoop DW Environment (Feasibility Test)



We observe that storage I/O can be much reduced by customizing caching SW to Hadoop workload.
(Deployment of SSD caching in SKT's infrastructure in progress)

Thank you

Please visit us at Booth #107