JLUG2017

Issues and Directions for the Next Generation Shared File System - 2
- How SSD based storage should be used? -

Shinji Sumimoto, Ph.D.
Next Generation Technical Computing Unit
FUJITSU LIMITED

Outline of This Talk

- Exascale Storage Design (JLUG2016)

- SSD Characteristics for Local File System

- How SSD Based Storage should be used?

Exascale Storage Design

- From JLUG2016 Presentation

Fujitsu's FEFS Development towards Exascale

- **Fujitsu will continue to develop Lustre based FEFS to realize the next generation exascale systems.**
  - Needs continuous Lustre enhancements
- **FEFS already supports Exa-byte class file system size**
  - However, several issues to realize real Exascale file system

- **Topics**
  - Exascale File System Design
  - Exascale Storage Design

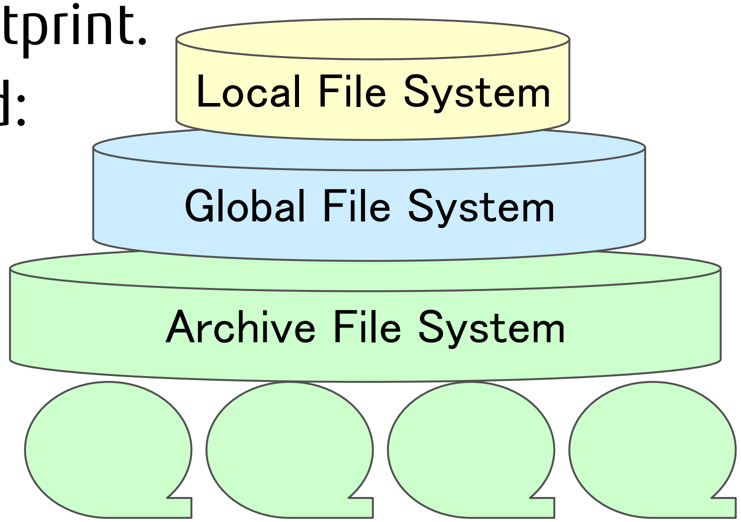
Exascale File System Design

K computer File System Design

- How should we realize High Speed and Redundancy together?
- How do we avoid I/O conflicts between Jobs?
- These are not realized in single file system.
  - Therefore, we have introduced Integrated Layered File System.

Exascale File System/Storage Design

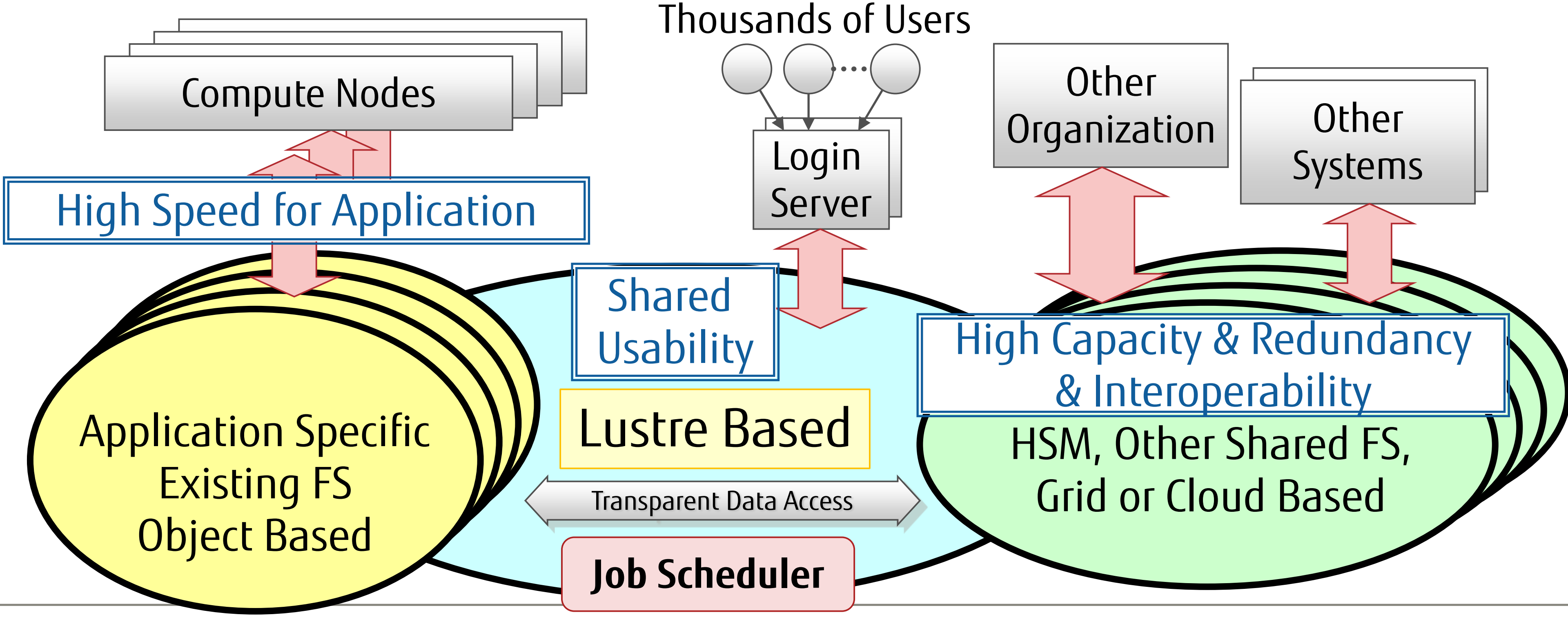
- Another trade off targets: Power, Capacity, Footprint
  - Difficult to realize single 1EB and 10TB/s class file system in limited power consumption and footprint.
- Third Storage layer for Capacity is needed: Three Layered File System
  - Local File System for Performance
  - Global File System for Easy to Use
  - Archive File System for Capacity



The Next Integrated Layered File System Architecture for Post-peta scale System (Feasibility Study 2012-2013)

FUJITSU

- Local File System o(10PB): Memory, SSD, HDD Based
  - Application Specific, Existing FS, Object Based, etc..
- **Global File System o(100PB): HDD Based**
  - Lustre Based, Ext[34], Object Based, Application Specific etc..
- Archive System o(1EB): HSM(Disk+Tape), Grid, Cloud Based
  - HSM, Lustre, other file system



Required Characteristics for the Next Integrated Layered File System

- **Application views:**
  - Local File System: Application Oriented File Accesses(Higher Meta&Data I/O)
  - Global File System: Transparent File Access
  - Archive System: In-direct Access or Transparent File Access(HSM)
- **Transparent File Access to the Global File System**
  - Local File System Capacity is not enough as much as locating whole data of Global File System
  - File Cache on node memory and Local File System enables to accelerate application performance

	Meta Perf.	Data BWs	Capacity	Scalability	Data Sharing in a Job	Data Sharing among Jobs
Local File System	◎	◎	×	◎	◎	×
Global File System	○	○	○	○	×	◎
Archive System	×	×	◎	×	×	×

This talk discusses about utilization of SSD for local file system

SSD Characteristics for Local File System

DRAM, SSD, HDD Memory Device Comparison

FUJITSU

- NAND Flash: Current SSD Devices
- PCM: Intel Optane

Table 1: Comparison of memory technologies.

	DRAM	PCM	NAND Flash	HDD
Read energy	0.8 J/GB	1 J/GB	1.5 J/GB [28]	65 J/GB
Write energy	1.2 J/GB	6 J/GB	17.5 J/GB [28]	65 J/GB
Idle power	~100 mW/GB	~1 mW/GB	1–10 mW/GB	~10 W/TB
Endurance	∞	$10^6 - 10^8$	$10^4 - 10^5$	∞
Page size	64B	64B	4KB	512B
Page read latency	20-50ns	~ 50ns	~ 25 μs	~ 5 ms
Page write latency	20-50ns	~ 1 μs	~ 500 μs	~ 5 ms
Write bandwidth	~GB/s per die	50-100 MB/s per die	5-40 MB/s per die	~200MB/s per drive
Erase latency	N/A	N/A	~ 2 ms	N/A
Density	1×	2 – 4×	4×	N/A

Note: The table contents are based mainly on [10, 15, 22].

CIDR 2011January 9–12, 2011 Asilomar, California http://cidrdb.org/cidr2011/Papers/CIDR11_Paper3.pdf

Rethinking Database Algorithms for Phase Change Memory
Shimin Chen, Phillip B. Gibbons Intel Labs Pittsburgh
and Suman Nath Microsoft Research

Endurance of PCM is 10-1000 times better than NAND Flash

Enterprise SSDs or Consumer SSDs

FUJITSU

	Enterprise Products		Consumer Products				
	Intel P3700	Intel P3608	Intel 750	Intel 600p	Samsung 950 pro	Samsung 960 Pro	Samsung 960 EVO
Capacity	800GB	1.6TB	1.2TB	1TB	512GB	1TB	1TB
Read Perf.	2.8GB/s	5.0GB/s	2.4GB/s	1.8GB/s	2.5GB/s	3.5GB/s	3.2GB/s
Write Perf.	1.9GB/s	2.0GB/s	1.2GB/s	0.6GB/s	1.5GB/s	2.1GB/s	1.9GB/s
Warranty	5 years	5 years	5 years	5 years	5 years	5 years	3 years
MTBF	2.0M	1.0M	1.2M	1.6M	1.5M	1.5M	1.5M
AFR	0.44%	0.87%	0.73%	0.54%	0.58%	0.58%	0.58%
DWPD	8TB/Day	4.8TB/Day	70GB/Day	40GB/Day	210GB/Day	430GB/Day	360GB/Day

https://www.intel.com/content/www/us/en/products/memory-storage/solid-state-drives.html
http://www.samsung.com/semiconductor/minisite/jp/ssd/consumer/overview.html

- **Same Level in Performance**
- **Differences in:**
  - DWPD(Data Writes per Day)
  - MTBF and AFR
- **Prices are increasing in proportion of their amount of flash cells**
  - Enterprise SSDs consist of flash cells as enough as their performance and endurance

Specification Difference in Intel P3700 Series

FUJITSU

	Enterprise Products			
	Intel P3700			
Capacity	400GB	800GB	1600GB	2000GB
Read Perf.	2.7GB/s	2.8GB/s	2.8GB/s	2.8GB/s
Write Perf.	1.1GB/s	1.9GB/s	1.9GB/s	1.9GB/s
Warranty	5 years	5 years	5 years	5 years
MTBF	2.0M	2.0M	2.0M	2.0M
AFR	0.44%	0.44%	0.44%	0.4%
DWPD	4TB/Day	8TB/Day	24TB/Day	34TB/Day

https://www.intel.com/content/www/us/en/products/memory-storage/solid-state-drives/data-center-ssds/dc-p3700-series.html

- DWPD increases in proportion of increasing its capacity

How about Intel Optane Products?

FUJITSU

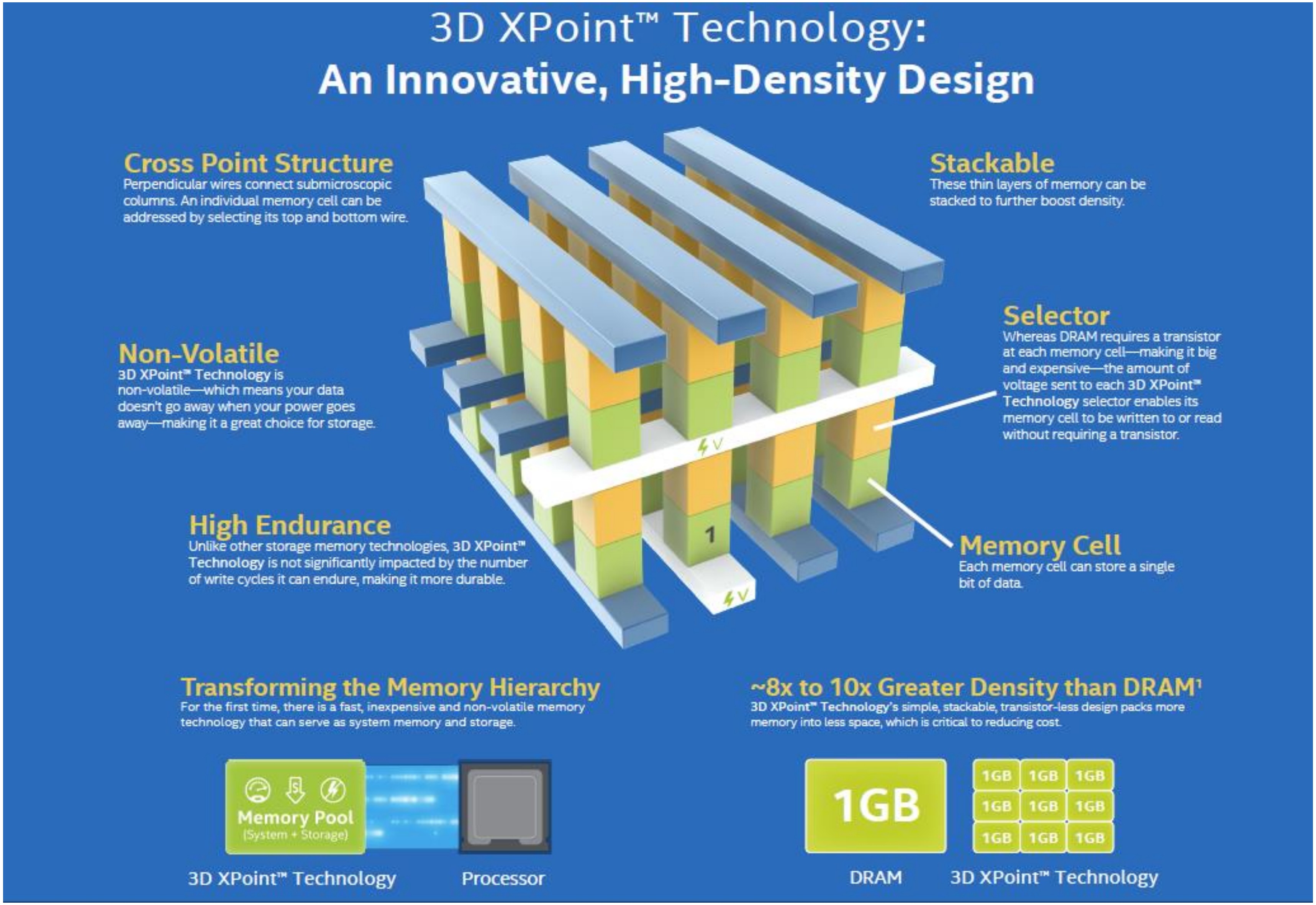
	Enterprise Products					Enthusiast
	Intel P3700	Intel P3608	Intel P4600	Intel P4500	Intel Optane P4800X	Intel Optane 900P
Capacity	800GB	1.6TB	1.6TB	1TB	375GB	480GB
Read Perf.	2.7GB/s	5.0GB/s	3.3GB/s	3.3GB/s	2.4GB/s	2.5GB/s
Write Perf.	1.9GB/s	2.0GB/s	1.4GB/s	0.6GB/s	2.0GB/s	2.0GB/s
K IOPS(R/W)	460/90	850/150	587/184	394/32	550/500	550/500
Latency(R/W)	20/20us	20/20us	79/34us	80/29us	10/10 us	10/10us
Warranty	5 years	5 years	5 years	5 years	5 years	5 years
MTBF	2.0M	1.0M	2.0M	2.0M	2.0M	1.6M
AFR	0.44%	0.87%	0.44%	0.44%	0.44%	0.54%
DWPD	8TB/Day	4.8TB/Day	4.7TB/Day	0.72TB/Day	11.2TB/Day	4.7TB/Day

https://www.intel.com/content/www/us/en/products/memory-storage/solid-state-drives/data-center-ssds.html

- **Intel Optane:**
  - Write IOPs is 2.7 times higher than that of P4600, but 375GB capacity is too small to use
  - DWPD 11.2TB/Day is not higher than expected, (3 times better than P3700/800G) but actual number of cells should be investigated.
  - Current cost is 30% higher than that of P3700 800GB (Amazon.com)

3D Xpoint

https://www.intelsalestraining.com/infographics/memory/3DXPointc.pdf



From the slides of Flash Memory Summit 2017

https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2017/20170808_FR12_Choe.pdf

Flash Memory Summit

XPoint Memory Overview

Tech Insights

- 16GB single die in a PKG
- Memory efficiency: 91.4 %
- Memory density (/Die): 0.62 Gb/mm^2
- Memory density (/Array): 0.69 Gb/mm^2

Top Metal View

Bpoly Level View

Package dimensions	18.0 mm x 14.0 mm x 1.10 mm thick
Manufacturer, part number, downstream	Intel, MEMPEK1W016GAXT, Optane™ 16GB memory module
Wafer size, foundry, process type	300 mm, Intel, 3D XPoint memory cell over CMOS
Die markings	<Intel logo> S15C (M) © 2014
Die size (from die seal)	16.16 mm x 12.78 mm (206.5 mm^2)
Die thickness	220 µm
Number, type of metals	5, 4 Cu and 1 Al and W used as word and bit lines
Minimum observed contacted logic gate pitch	0.38 µm
Minimum observed logic transistor gate length	0.086 µm
Minimum metal pitch	84 nm
3D XPoint memory bit line (word line) pitch	38.5 nm
3D XPoint memory word line (bit line) pitch	40 nm
Memory cell area	0.0015 μm^2
Technology generation	20 nm
Feature measured to determine process generation	Half bit line (word line) pitch

Flash Memory Summit 2017
Santa Clara, CA

5

FUJITSU

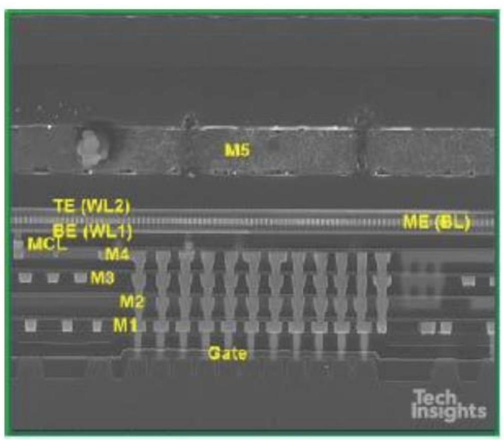
https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2017/20170808_FR12_Choe.pdf

Flash Memory Summit

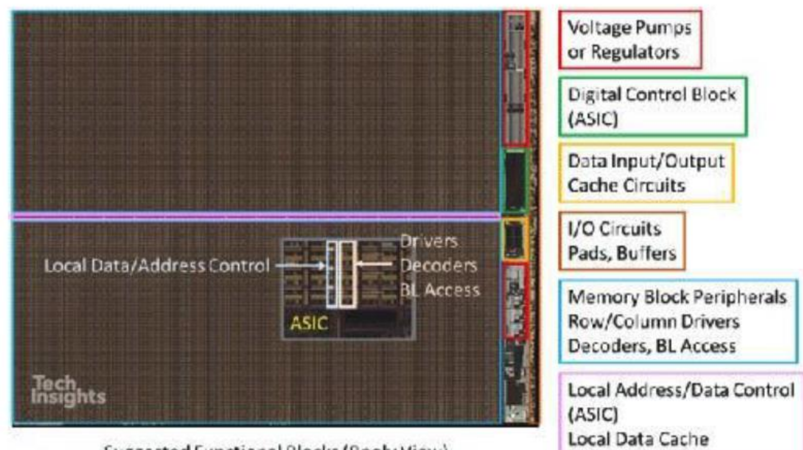
XPoint: Process Integration

TechInsights

- GST-based PCM (Phase Change Memory) between M4 and M5
- Storage layer vertically stacked on Selector
- Se-Ge-Si ternary phased OTS Selector with As doped
- Double memory cell stacked
- 1 Poly Si (Co-silicide), 5 Metals (excluding memory/WL/BL layers)



SEM X-Section (Array)



Suggested Functional Blocks (Bpoly View)

Flash Memory Summit 2017
Santa Clara, CA

6

From the slides of Flash Memory Summit 2017

FUJITSU

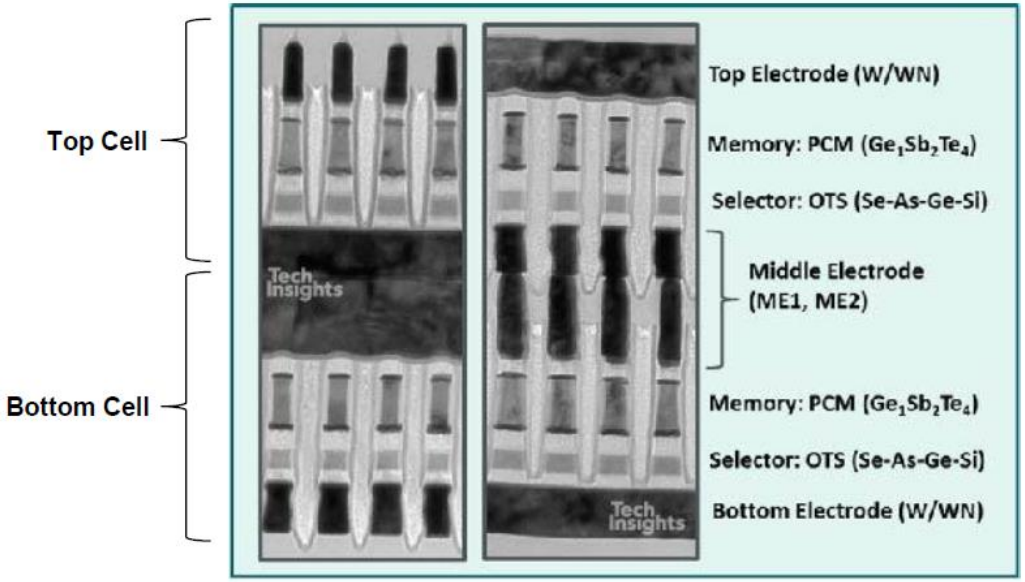
https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2017/20170808_FR12_Choe.pdf

Flash Memory Summit

XPoint: Memory/OTS Elements

Tech Insights

- **Top & bottom cell stacked**
- **TWL/TE/PCM/ME/OTS/BE/BL2/BL1/TE/PCM/ME/OTS/BE/BWL**
- **PCM:** $Ge_{0.12}Sb_{0.29}Te_{0.54}(Si_{0.05})$, **OTS:** $Se_{0.44}As_{0.29}Ge_{0.1}Si_{0.17}$,



Flash Memory Summit 2017
Santa Clara, CA

7

From the slides of Flash Memory Summit 2017

https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2017/20170808_FR12_Choe.pdf

XPoint, could be ……

- ✓ 1,000 times faster than NAND Flash
- ✓ 10 times denser than DRAM → Really?
- ✓ 1,000 times better endurance than NAND

vs. 3D NAND?

Flash Memory Summit 2017
Santa Clara, CA

8

Evaluation of Intel Optane

From the slides of Flash Memory Summit 2017

FUJITSU

https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2017/20170808_FR12_Choe.pdf

Flash Memory Summit

XPoint is ......

TechInsights

vs. DRAM

6 times denser than Micron 20 nm DRAM

3 times denser than Samsung 1x DRAM

vs. NAND

18% memory density of Toshiba/SanDisk 64L NAND

Higher memory cell area efficiency than 2D NAND

Relatively lower cell area efficiency than 3D NAND

Flash Memory Summit 2017
Santa Clara, CA

15

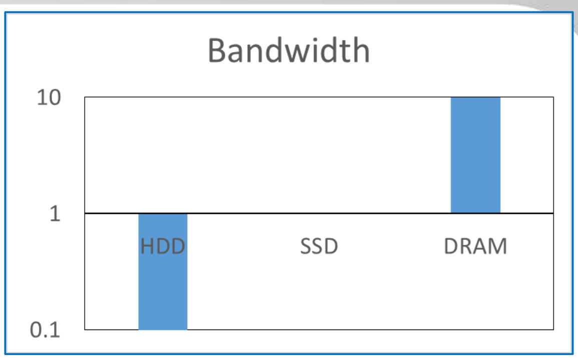
- vs. Intel P3700/800GB
  - Latency of XPoint is two times better
  - Endurance of XPoint is three times better

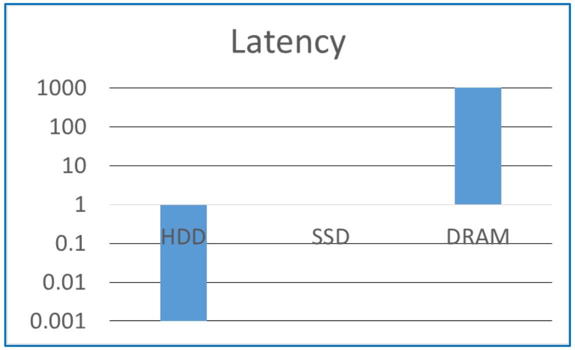
How SSD Based Storage should be used?

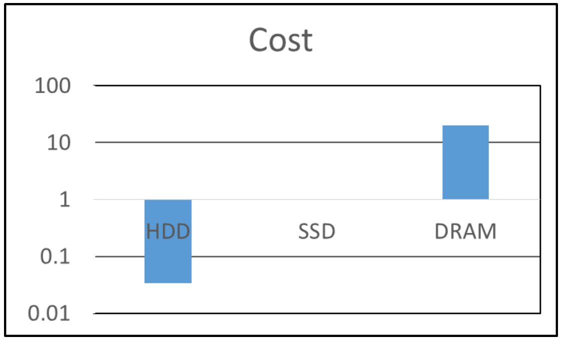
Utilizing SSD based storage

FUJITSU

- **Characteristics of SSD:**
  - Bandwidth Performance:
    - vs. HDD: - 10 times faster,
    - vs. DRAM(DIMM): - 10 times slower
  - Latency:
    - vs. HDD: -1000 times faster
    - vs. DRAM(DIMM): -1000 times slower
  - Capacity per cost (amazon.com price):
    - vs. HDD: 30- times higher
    - vs. DRAM(DIMM): -20 times lower
  - Endurance:
    - Limited lifetime writes
- **To utilize SSD characteristics:**
  - Reduction of HDD access
  - Lifetime write control: Elimination of useless writes
- **Whether useless or not depends on file I/O access pattern**
  - Needs to investigate file I/O usage on applications



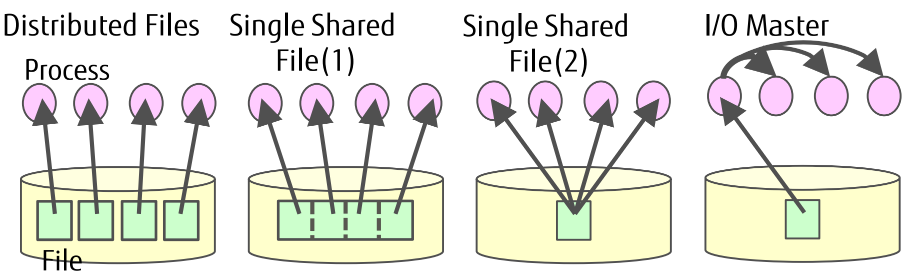




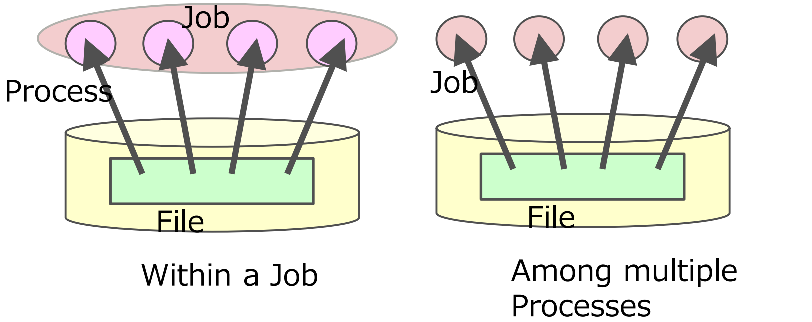
Three Scopes of File I/O Usages on Applications

- **File Lifetime:**
  - Persistent Files: Input Files, Output Files
  - Temporary Files: Input Files, Output Files

- **Access Pattern:**
  - Distributed Files: for each process
  - Single Shared File : partial access, concentrate access to same data
  - Master-slave: Master does whole File I/O



- **Data Sharing:**
  - Within a job
  - Among multiple jobs(under designing)

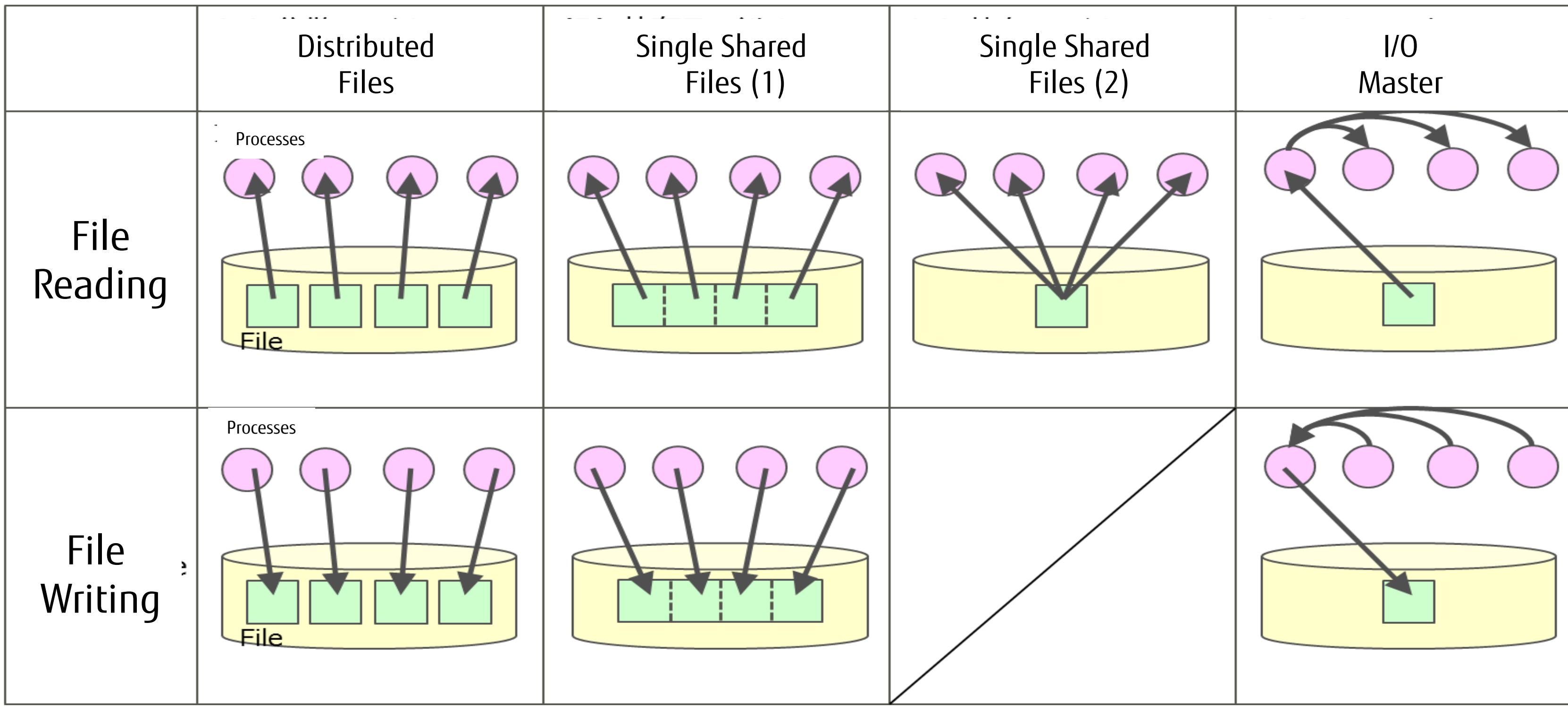


File Lifetime for effective SSD use

- Persistent files in a job are located on SSD as file cache
  - SSD based storage capacity is smaller than that of the global FS
  - Asynchronous data transfer is effective between the local and global FS
- Temporary files in a job should be located on SSD to eliminate the global FS accesses
- But, how persistent file cache on SSD should be used?
  - It depends on file access patterns

Application's Access Pattern and SSD Cache Effects

■ Comparison of Effective Pattern for SSD based storage

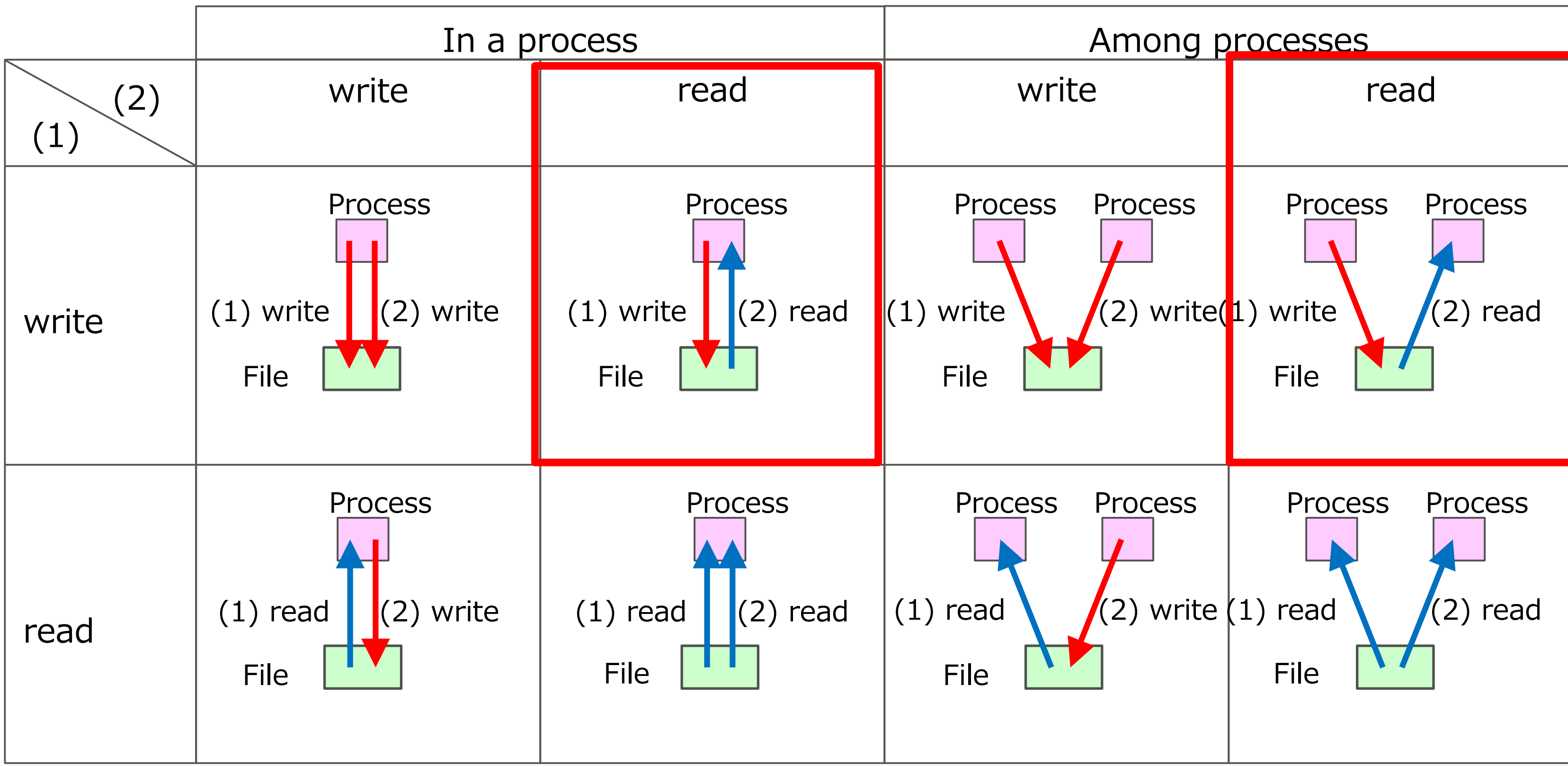


File Read: Effects	Rereading Case: ◎ Non Rereading : ×	Rereading Case: ◎ Non Rereading : ×	Rereading Case: ◎ Non Rereading : ×	Rereading Case: ◎ Non Rereading : ×
File Write: Effects	Rewriting Case: ◎ Non Rewriting : O	Rewriting Case: ◎ Non Rewriting : O		Rewriting Case: ◎ Non Rewriting : O

Data Sharing in a Job on SSD

FUJITSU

- Write-Read in a process and among processes are effective to use SSD
- For Persistent Files: File cache of global file system should be shared among processes
- For Temporary Files: Two types of temporary file systems are effective to use SSD
  - Temporary Local System (in a process)
  - Temporary Shared File System (among processes)



Data Sharing among multiple jobs on SSD

- Write-Read among multiple jobs are effective
- Issues:
  - Local File System Data Lifetime management
    - When file data will be removed from SSD?
  - How to realize SSD capacity management
    - With relation with Job scheduler or not
  - Performance and Availability
- Needs to be designed how to share file on global file system and local file system

How SSD based storage should be used?

FUJITSU

Life Time

- Persistent files in a job are located on SSD as file cache
- Temporary files in a job should be located on SSD to eliminate the global FS accesses

Application's Access Pattern

- Non reusable file in file reading should not use SSD based storage

Data Sharing in a Job

- Write-Read in a process and among processes are effective to use SSD
- For Persistent Files: File cache of global file system should be shared among processes
- For Temporary Files: Two types of temporary file systems are effective to use SSD
  - Temporary Local System (in a process)
  - Temporary Shared File System (among processes)

Data Sharing among multiple jobs

- Write-Read among multiple jobs are effective to use SSD
- Needs to be designed how to share file cache on global and local file system

SSD lifetime writes(DWPD) Issue

- SSD whose DWPD is higher than that of daily use will be a choice shaping tomorrow with you