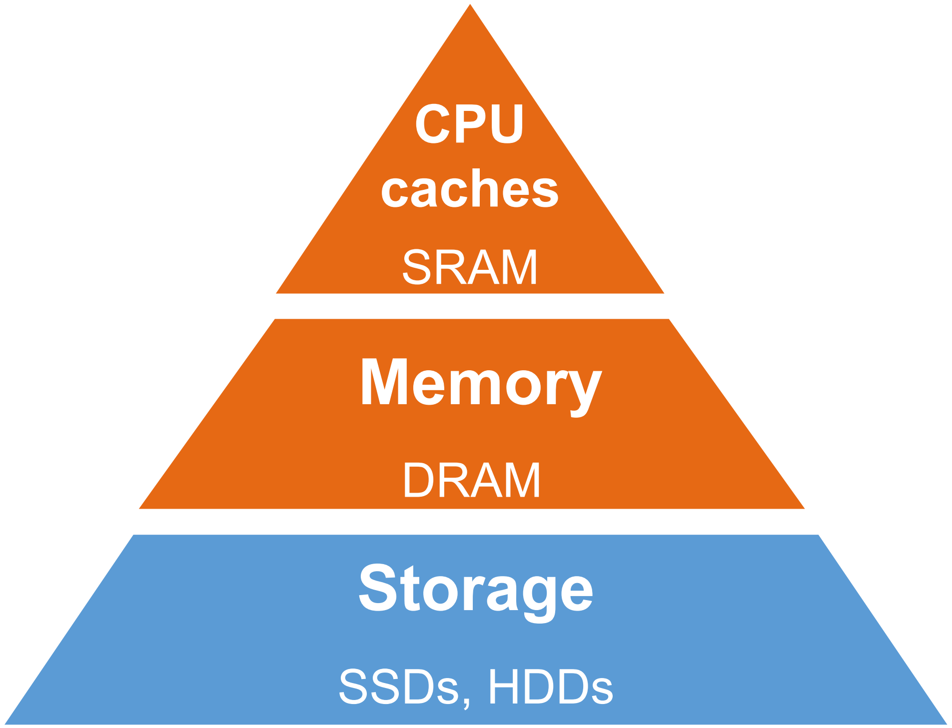
Rethinking the Performance/Cost of Persistent Memory and SSDs

Kaisong Huang Darien Imai Tianzheng Wang Dong Xie

SFU SIMON FRASER UNIVERSITY

PennState

The Storage Hierarchy as We Knew It



Layers with clear boundaries
Memory: fast but volatile
Storage: slower than memory but persistent

Caching stores hugely successful

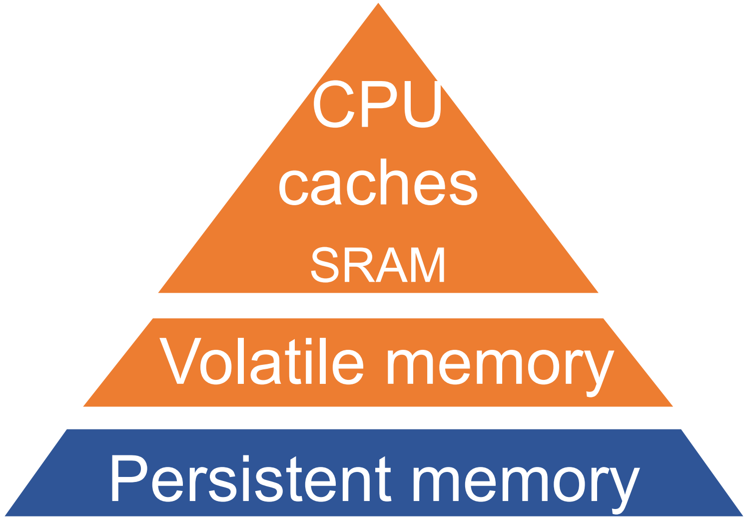
- Hot data in buffer pool (DRAM)
- The whole dataset on drives
- Practical & Cost-effective

…is being disrupted by two trends

Trend 1: (Persistent) Memory Meets Persistence

Persistent memory, generally speaking
- Byte addressable
- Persistence
- Large capacity
- Cheaper than DRAM

Intel Optane Persistent Memory 200 (3D XPoint)
- Peak read: 7.4 GB/s per DIMM
- Peak write: 2.3 GB/s per DIMM
- Capacity: 128/256/512 GB per DIMM



“PM camp”

(a lot of attention)

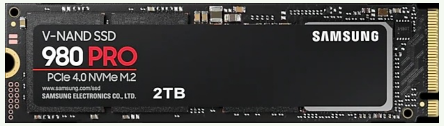
Buffer pool + SSD

Single-level index/store

“SSDs no more, cheaper than DRAM – all in!”

Trend 2: SSD Approaches (Persistent) Memory

- New materials
  - 3D V-NAND Flash or 3D XPoint
- New interconnection
  - PCIe Gen4
- New software stack
  - SPDK, io_uring

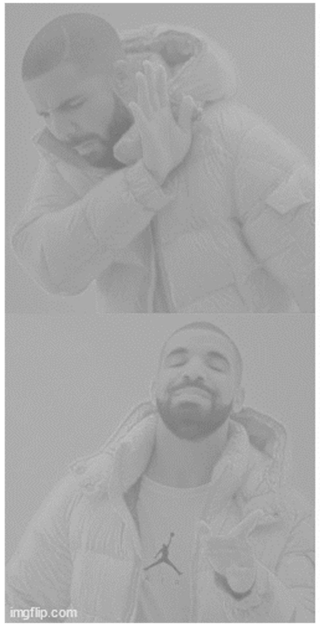


Intel Optane DC SSD P5800X
- Peak read: 7.4 GB/s
- Peak write: 7.4 GB/s
- Capacity: 400/800/1600GB x # drives

vs.

Intel Optane PMem 200 (128GB DIMM)
- Peak read: 7.4 GB/s
- Peak write: 2.3 GB/s
- Capacity: 128GB x # memory channels

"PM camp"
(a lot of attention)



Buffer pool + SSD

Single-level index/store

"SSD camp"
(relatively quieter)

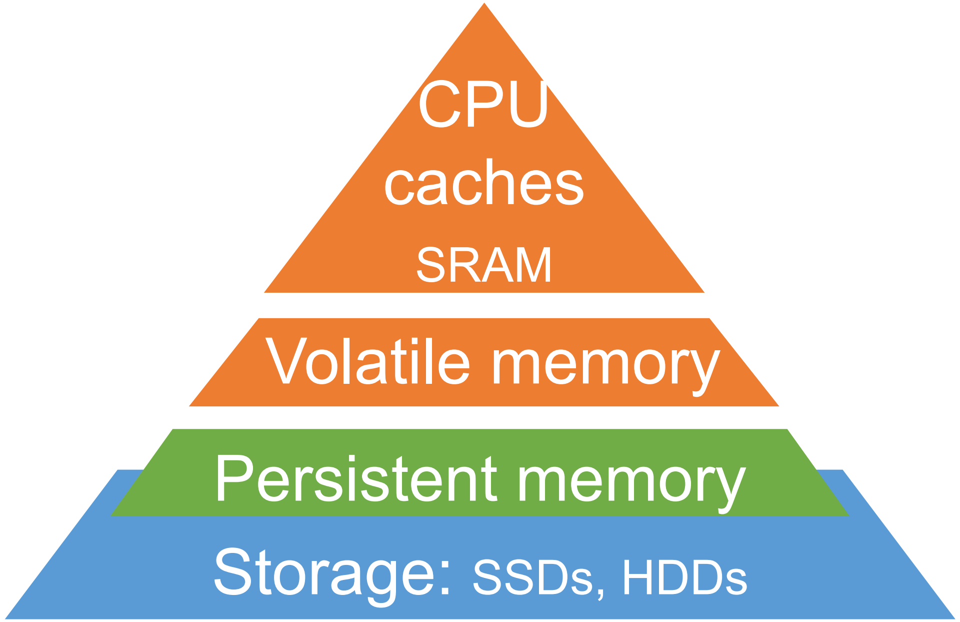
Single-level index/store

~In-memory performance atop SSD

With faster SSDs, match or outperform PM indexes?

SFU SIMON FRASER UNIVERSITY PennState

The Storage Jungle



Layers with overlapping properties

Memory not necessarily volatile

Storage not necessarily slower than memory

PM vs. SSD Servers: What to Consider

Rigid installation requirements

- Strict population rules
  - >= 1 DRAM DIMM per controller

→ Overprovisioning

- Clock down frequency

→ Affect overall memory performance

Non-trivial CPU cost

- Synchronous `load/store`

→ High-end CPU cores wasted

Flexible installation requirements

- DRAM requirement decoupled
- Few population rules (e.g., RAID)

→ Nothing overprovisioned

Low CPU cost

- Asynchronous DMA

→ Overlap I/O operations and computing

PM vs. SSD Servers: Costs

$ per GB	PM1 128G	PM6 768G	P4800X 375G
Storage-only	$4.27	$4.27	$2.66
Storage+DRAM	$13.32	$5.78	$5.75
Storage+DRAM+ CPU (minimum)	$18.23	$7.76	$5.92
Storage+DRAM+ CPU (full)	$32.98 (Total: $4,221.69)	$9.06 (Total: $6,955.44)	$12.46 (Total: $4,673.94)

Observations:

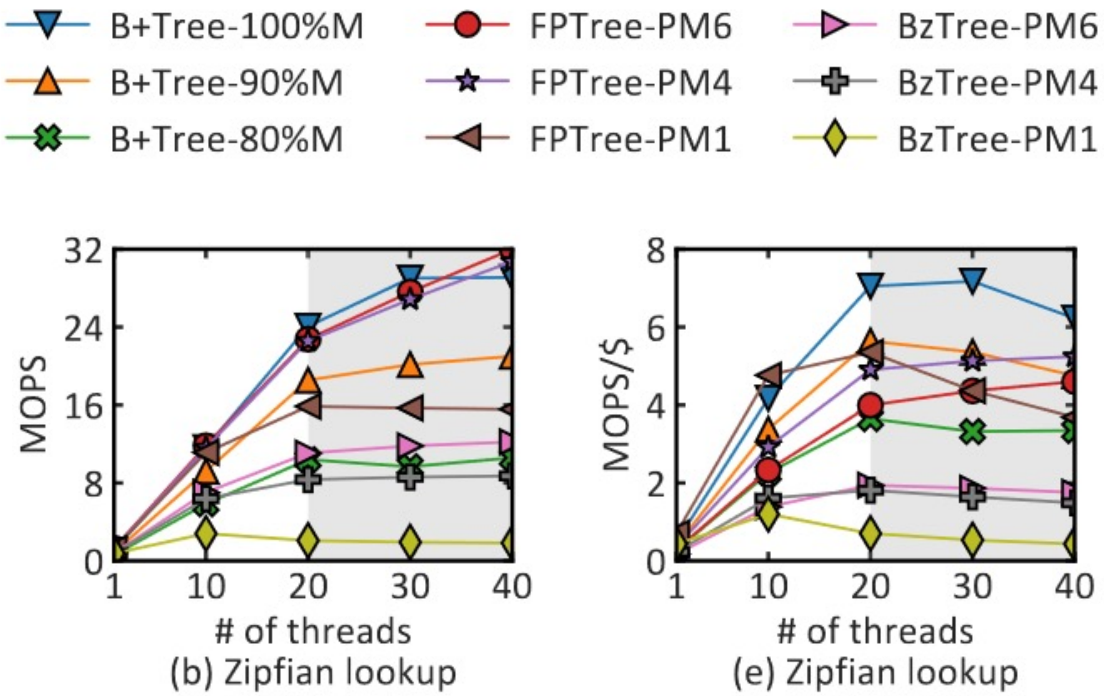
Same material, but PM is more expensive than SSD

DRAM significantly increases the unit prices

10 threads to saturate PM (write bandwidth), 1 thread to saturate SSD

Not fair ☹

PM vs. SSD Servers: Performance



(more details in paper)

FPTree & BzTree:

Tailor-made, optimized for PM

B+Tree:

Coursework-grade (!) atop P4800X

Takeaways:

- Memory-resident? Use SSD + buffer pool
- PM indexes still rely on DRAM to gain performance
- P4800X is very competitive with PM1

Final Thoughts

Before you invest in PM…

- PM hardware is still too expensive;
  - High-end CPU cores for “I/O” + extra DRAM costs
- PM software stack is also “expensive”
  - A steep learning curve, complex programming model

Is SSD a done deal? No.

- SSD is usually more cost-effective
  - Even with suboptimal implementation
- Explore newer storage interfaces (e.g., SPDK)

Full paper at CIDR 2022: SSDs Striking Back: The Storage Jungle and Its Implications on Persistent Indexes

Code: https://github.com/sfu-dis/ssd-vs-pm

Thank you!

SFU SIMON FRASER UNIVERSITY PennState