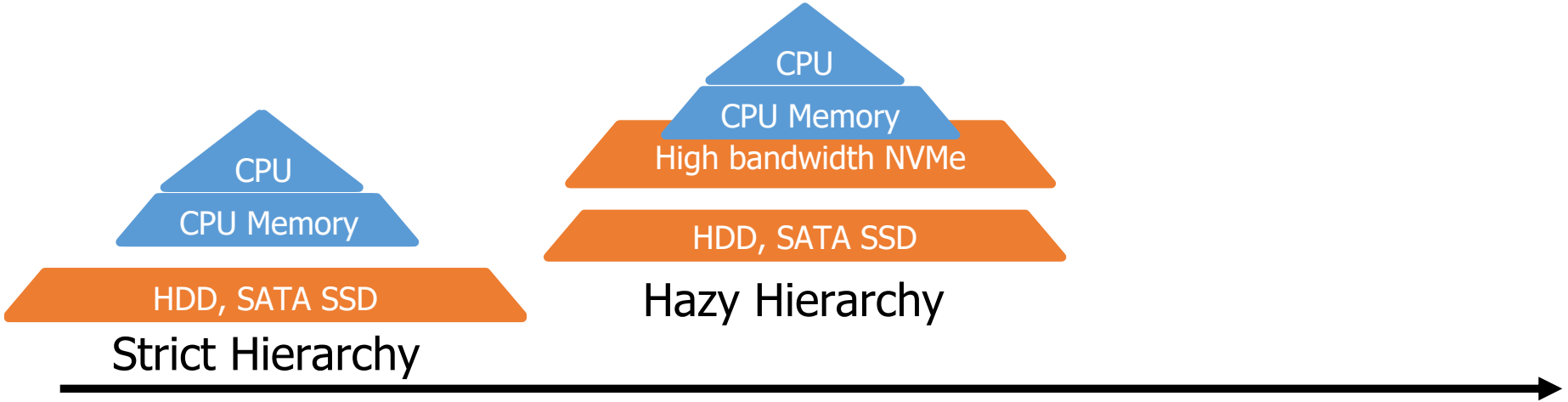
GPU-accelerated Queries on Flash Storage

Hamish Nicholson

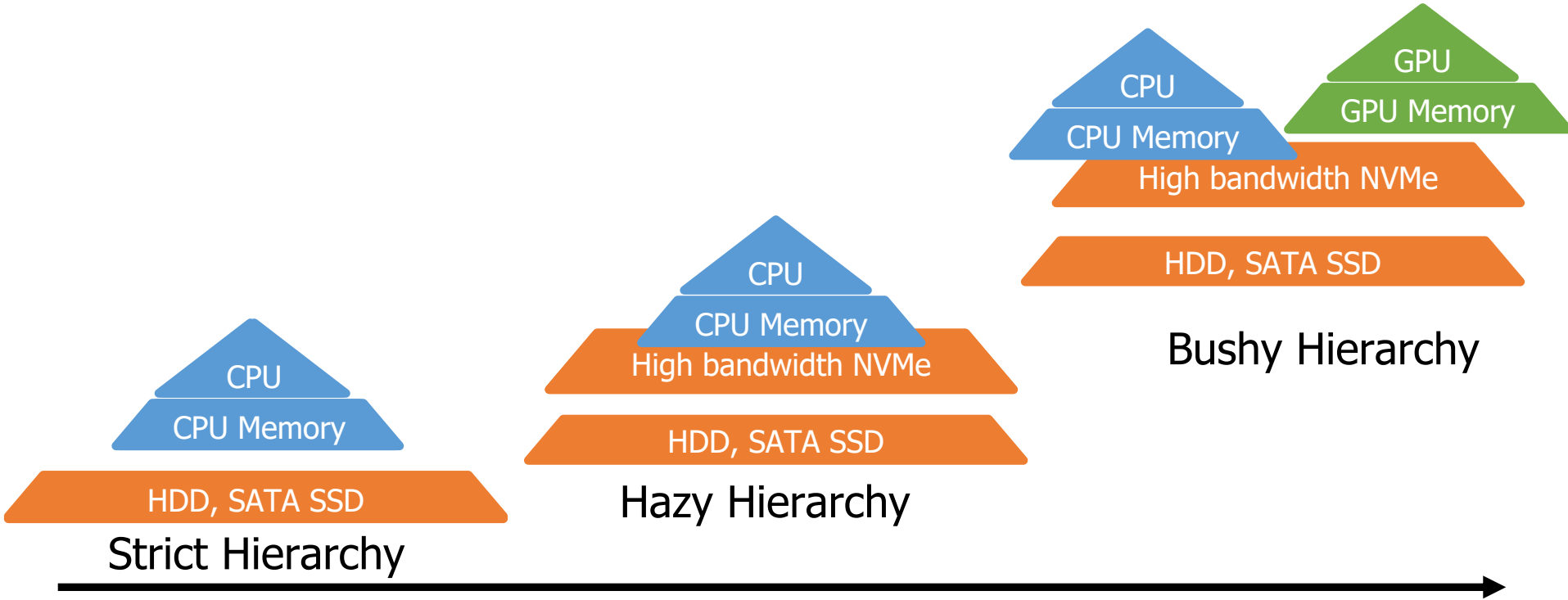
11.10.2022

Storage Anarchy: The Breakdown of the Strict Hierarchy

Storage Anarchy: The Breakdown of the Strict Hierarchy

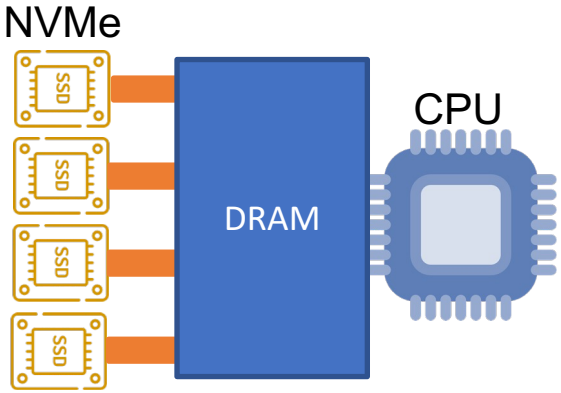


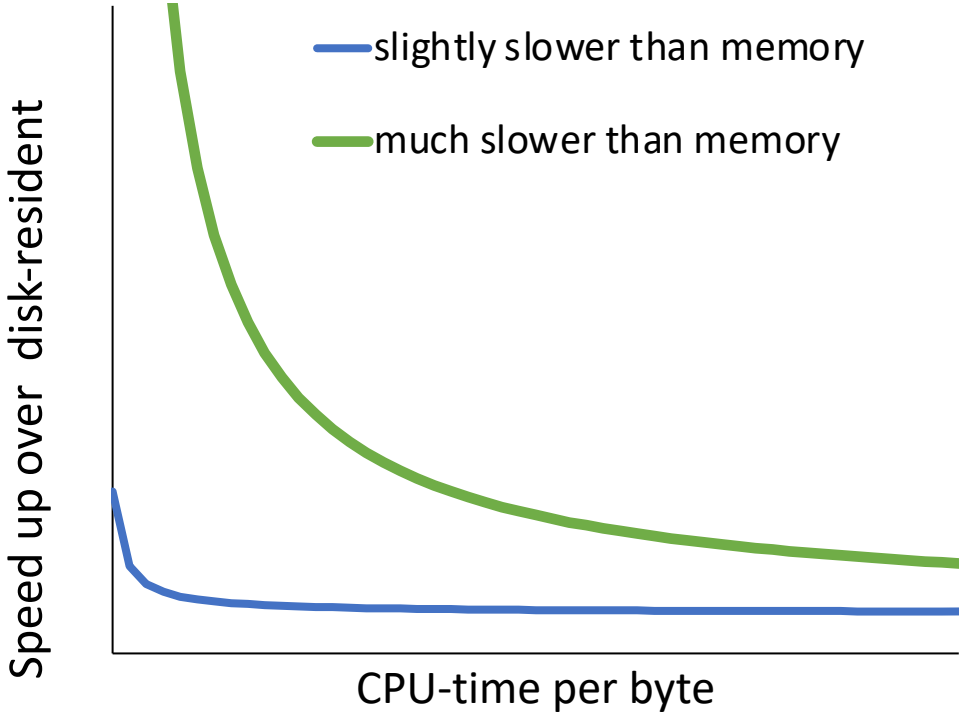
Storage Anarchy: The Breakdown of the Strict Hierarchy



Anarchy enables improved resource utilization ($$$)

Storage Bandwidth Approaches Memory Bandwidth

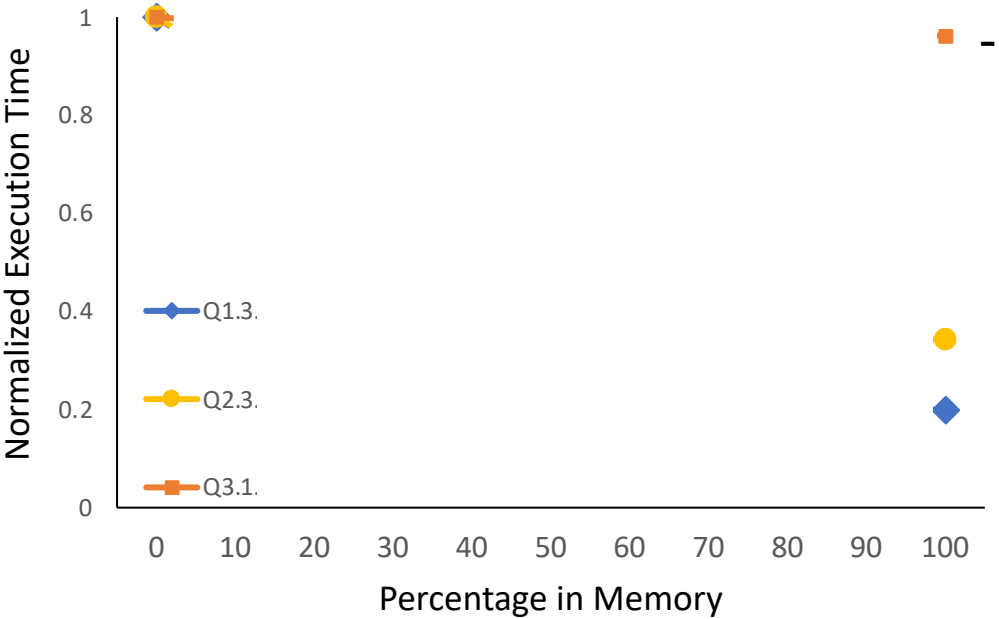




Faster storage = memory savings

Benefit of Memory is Workload Dependent

2x 12-core Xeon Skylake 5118
4x (NVMe, PCIe 3 x8)
SSB1000
96GB working set



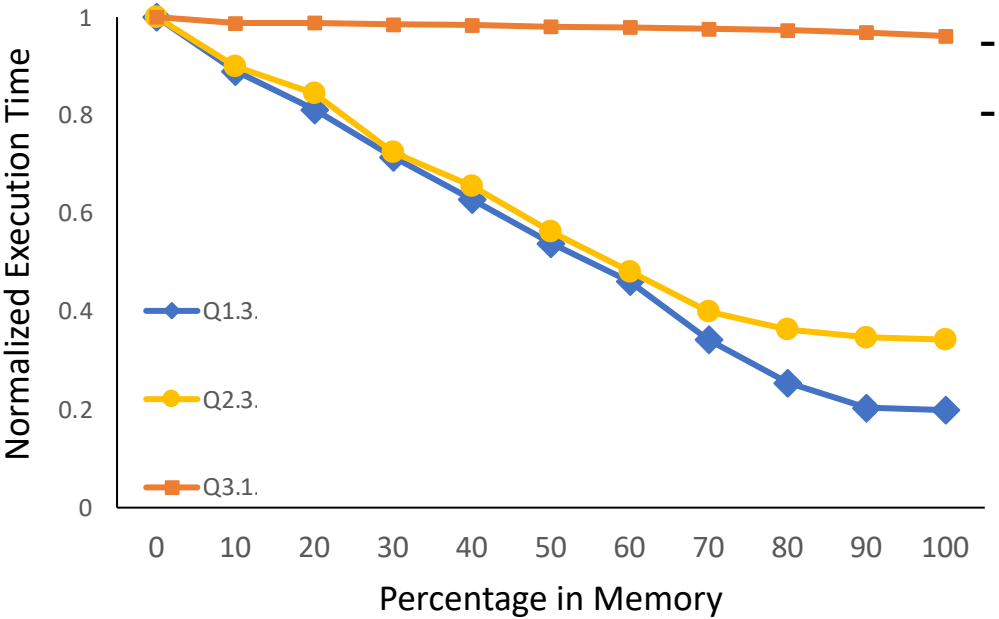
Benefit of memory is:

- Query dependent

Match storage bandwidth to workload throughput

Benefit of Memory is Workload Dependent

2x 12-core Xeon Skylake 5118
4x (NVMe, PCIe 3 x8)
SSB1000
96GB working set

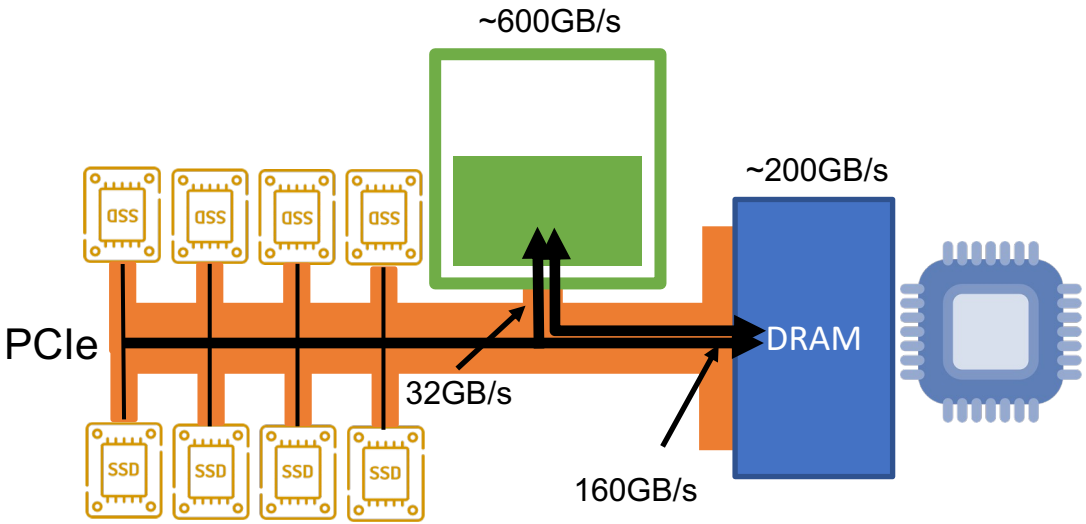


Benefit of memory is:

- Query dependent
- Has diminishing returns

Match storage bandwidth to workload throughput

Bushy Hierarchies



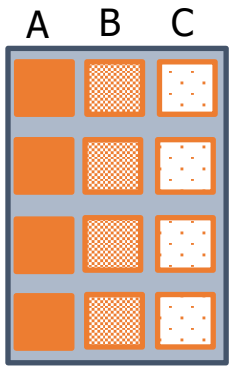
Heterogeneous:

- Compute
- Capacity
- Access

Data placement & access severely complicated

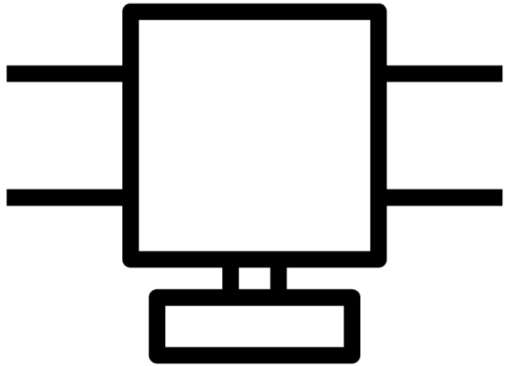
Direct Storage-to-GPU Access

NVMe Storage



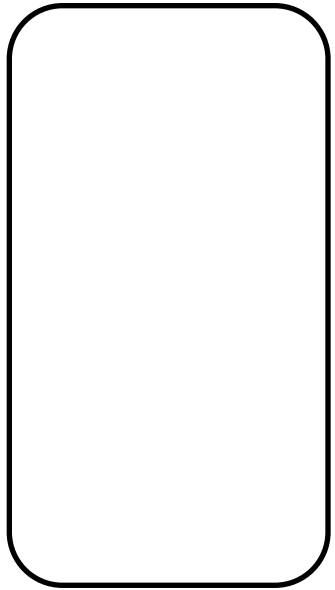
SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

T = 0



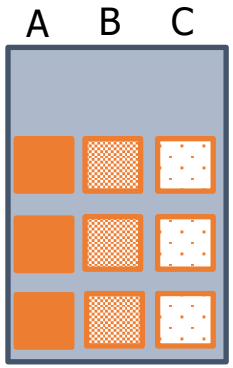
GPU Pipeline
3 block / T

Processed Data



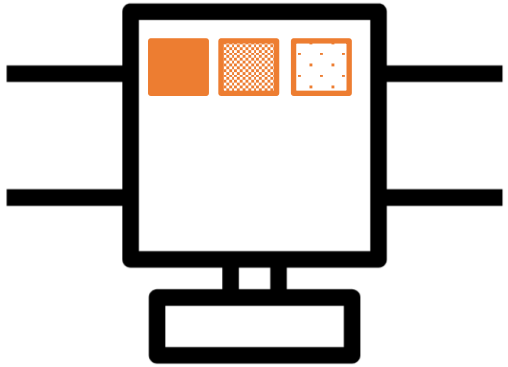
Direct Storage-to-GPU Access

NVMe Storage



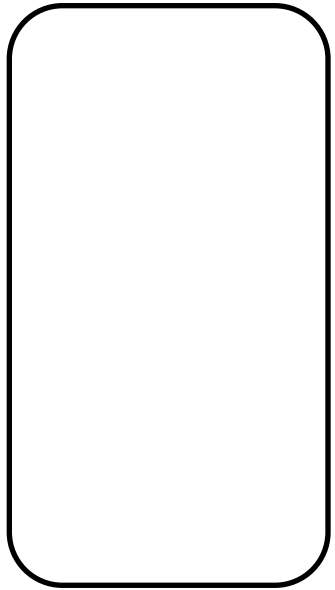
SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

T = 1



GPU Pipeline
3 block / T

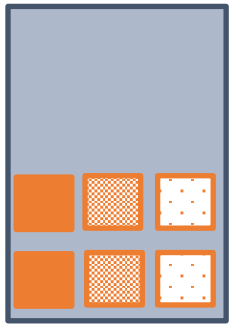
Processed Data



Direct Storage-to-GPU Access

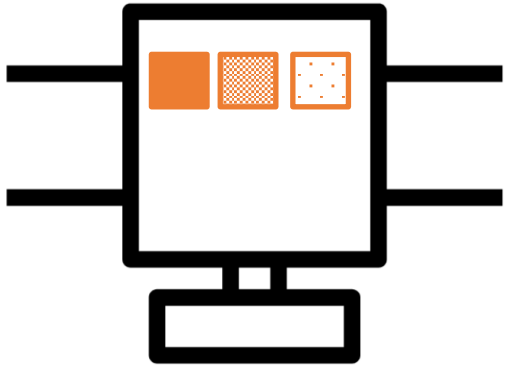
NVMe Storage

A B C



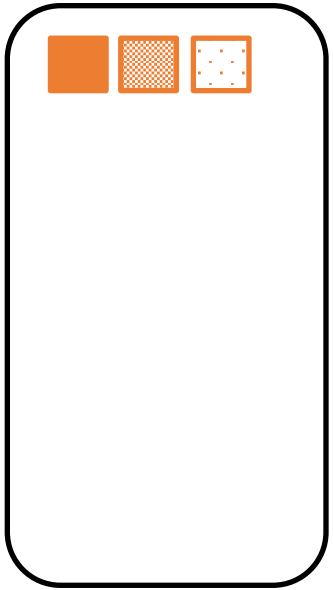
SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

T = 2



GPU Pipeline
3 block / T

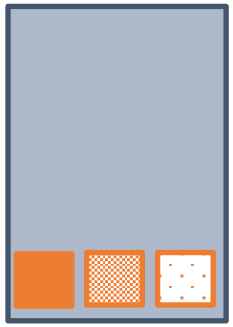
Processed Data



Direct Storage-to-GPU Access

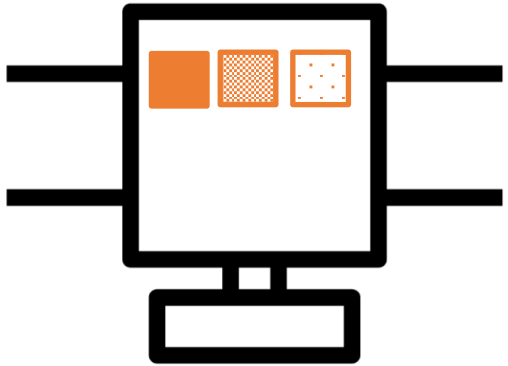
NVMe Storage

A B C



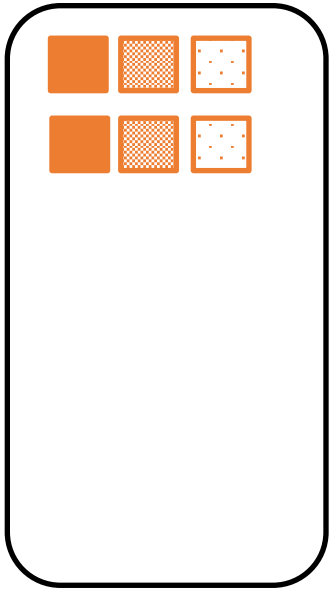
SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

T = 3



GPU Pipeline
3 block / T

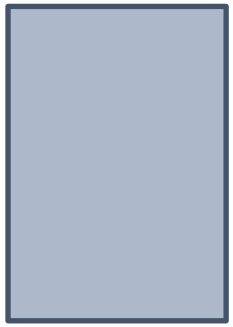
Processed Data



Direct Storage-to-GPU Access

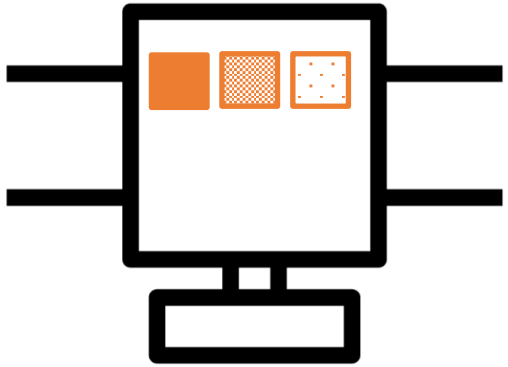
NVMe Storage

A B C



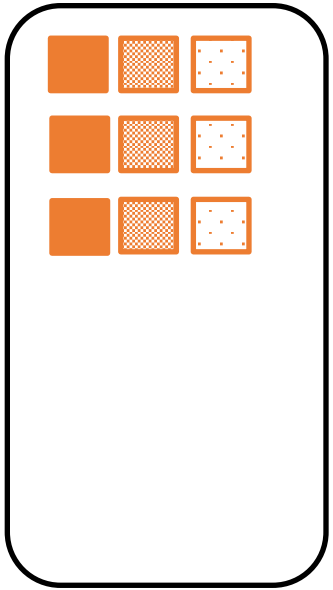
SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

T = 4



GPU Pipeline
3 block / T

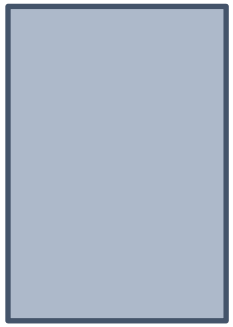
Processed Data



Direct Storage-to-GPU Access

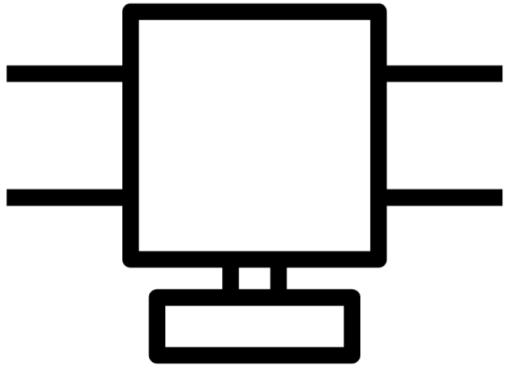
NVMe Storage

A B C



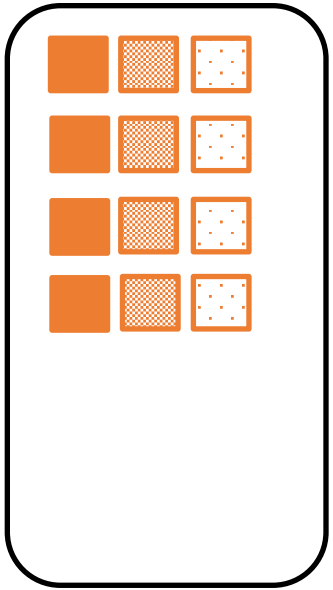
SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

T = 4



GPU Pipeline
3 block / T

Processed Data

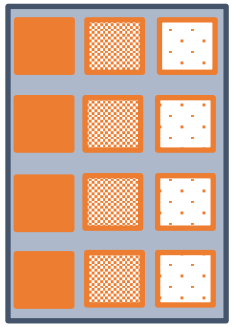


Limited by interconnect BW

System Memory as a Staging Ground

NVMe Storage

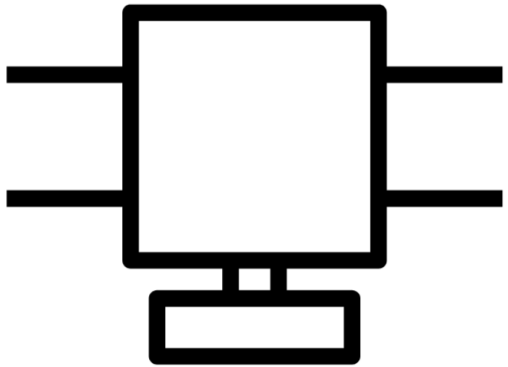
A B C



System Memory

SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

T = 0



GPU Pipeline

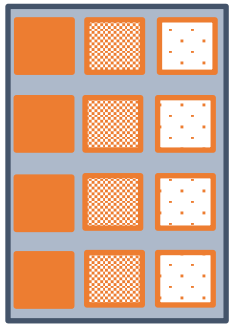
3 block / T

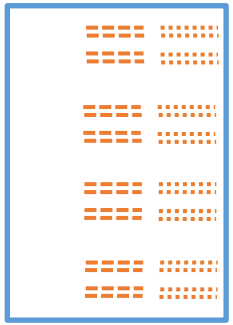
Processed Data

System Memory as a Staging Ground

NVMe Storage

A B C

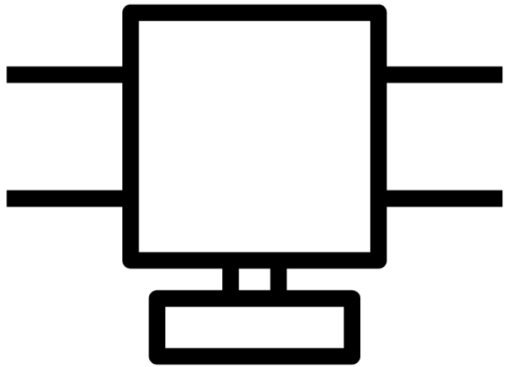




System Memory

SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

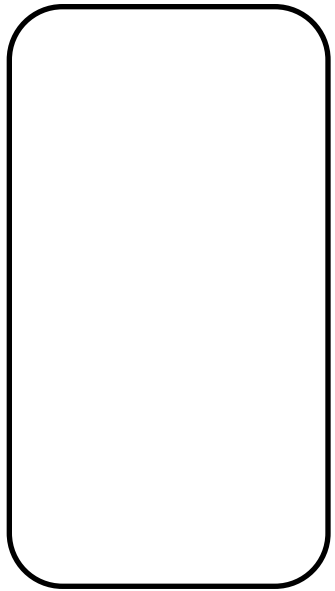
T = 0



GPU Pipeline

3 block / T

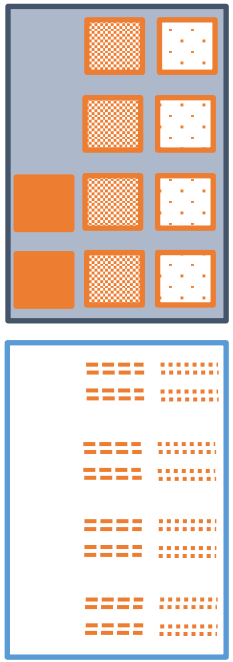
Processed Data



System Memory as a Staging Ground

NVMe Storage

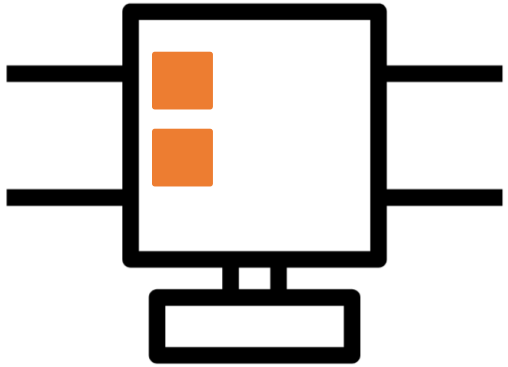
A B C



System Memory

SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

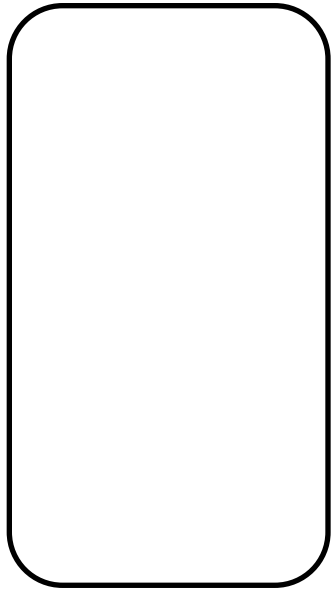
T = 1



GPU Pipeline

3 block / T

Processed Data

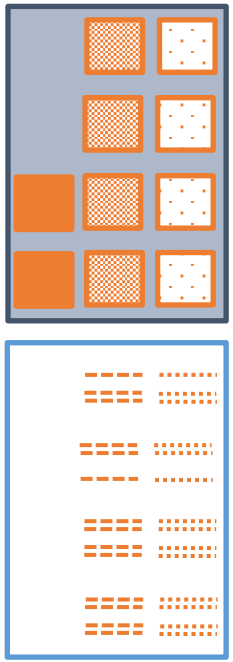


Leverage memory for fine grained accesses

System Memory as a Staging Ground

NVMe Storage

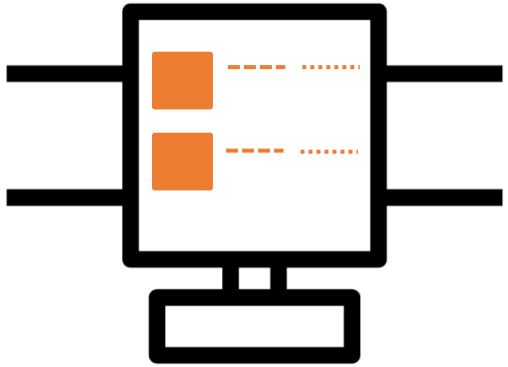
A B C



System Memory

SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

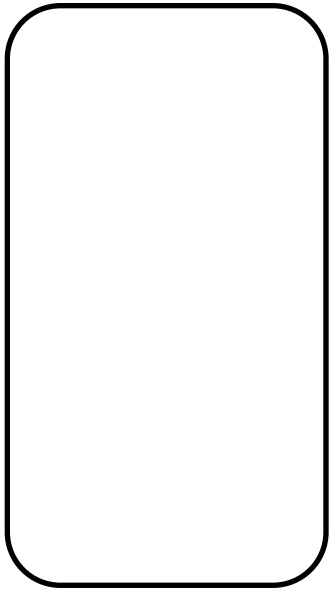
T = 1



GPU Pipeline

3 block / T

Processed Data

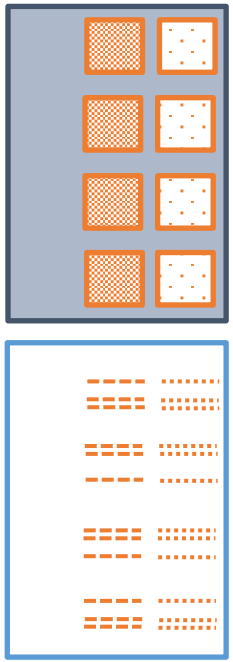


Leverage memory for fine grained accesses

System Memory as a Staging Ground

NVMe Storage

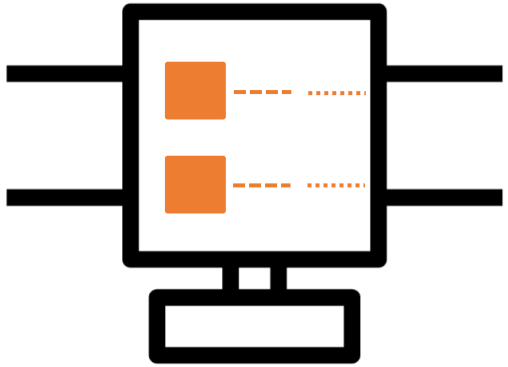
A B C



System Memory

SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

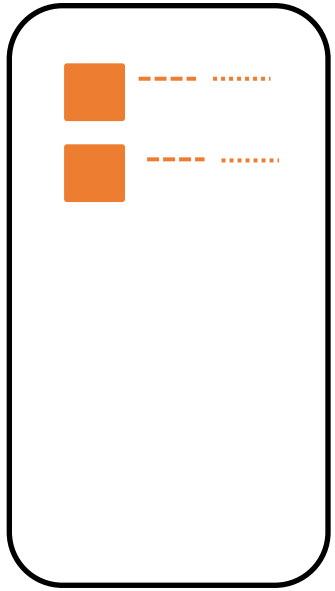
T = 2



GPU Pipeline

3 block / T

Processed Data

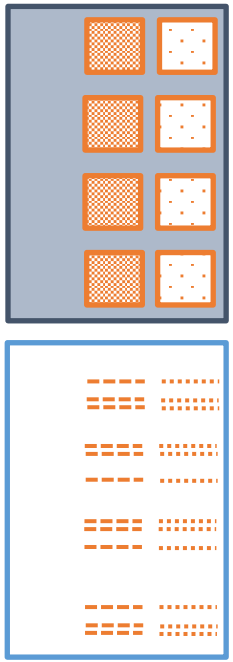


Leverage memory for fine grained accesses

System Memory as a Staging Ground

NVMe Storage

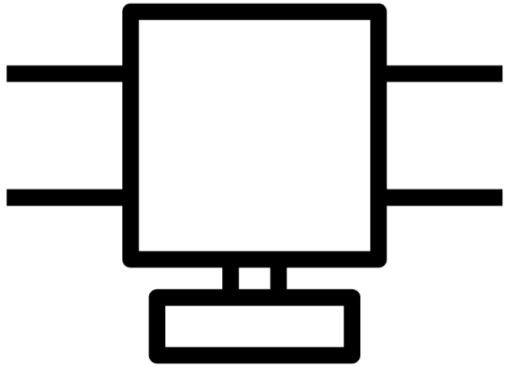
A B C



System Memory

SELECT T.c FROM T WHERE T.a < 50 AND T.b > 42

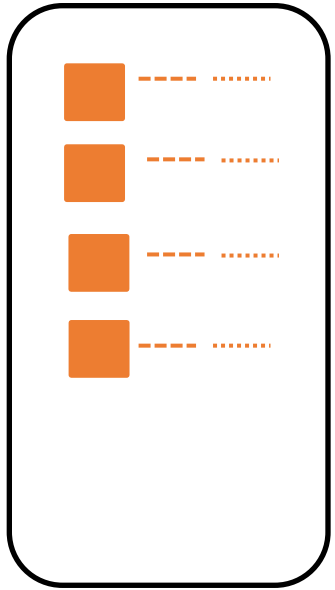
T = 2



GPU Pipeline

3 block / T

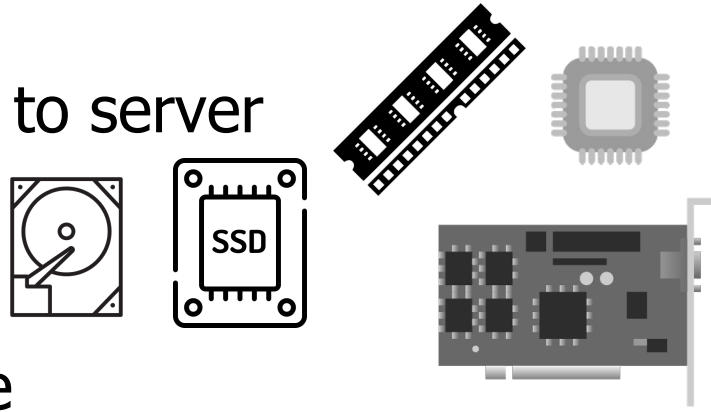
Processed Data



Leverage memory for fine grained accesses

Systems Must be Designed for the Anarchy

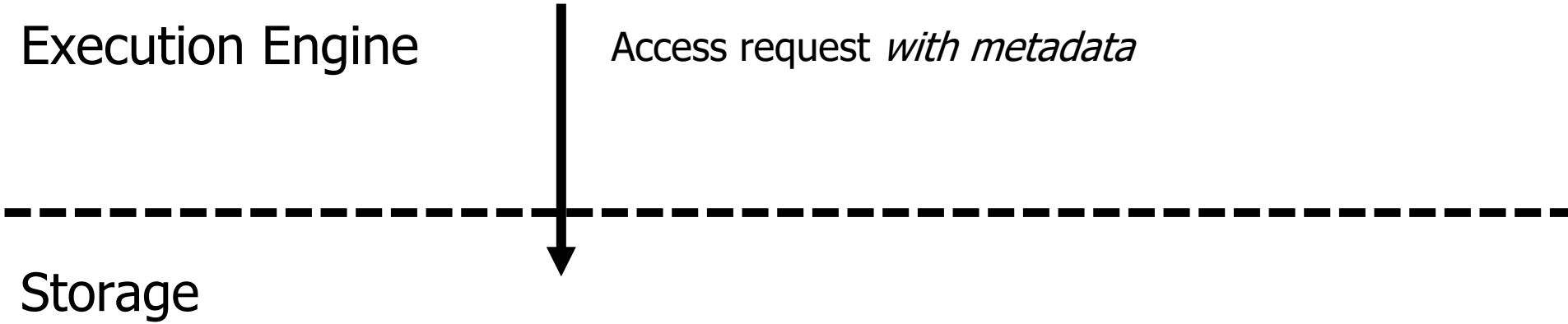
- Hardware topology varies from server to server



- Workload often only known at runtime

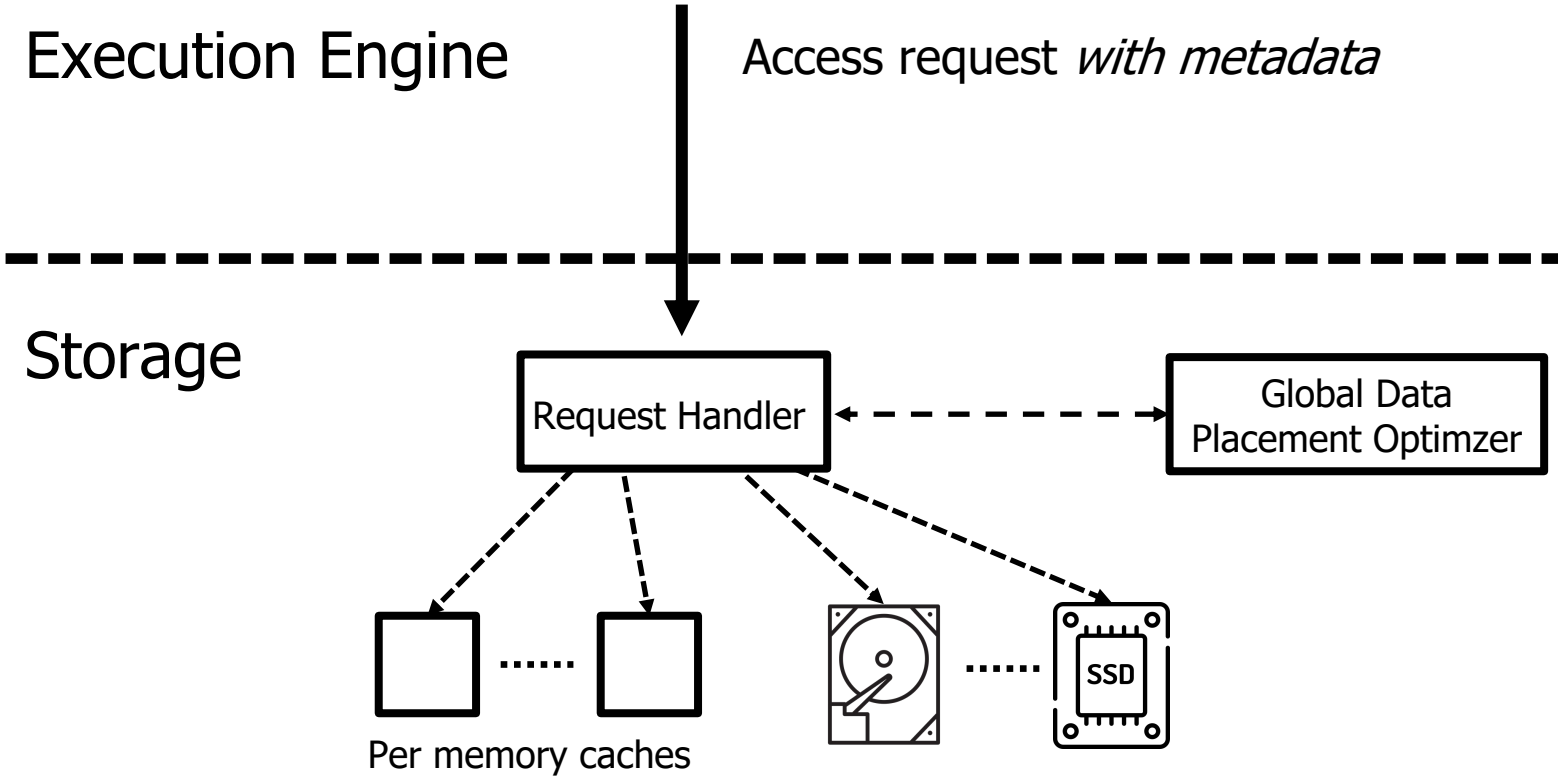
Storage engines need to adapt at runtime

The Anarchists Storage Engine



Topology provides options && workload guides decisions

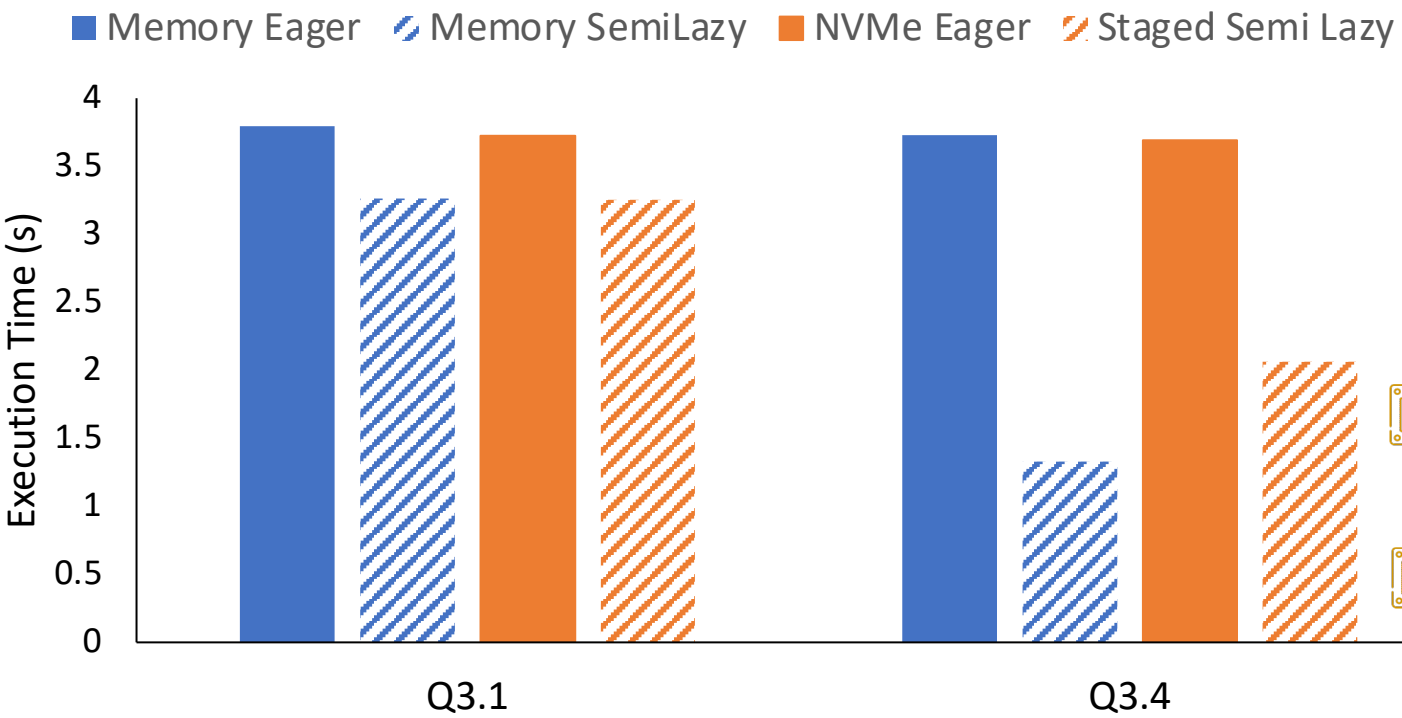
The Anarchists Storage Engine

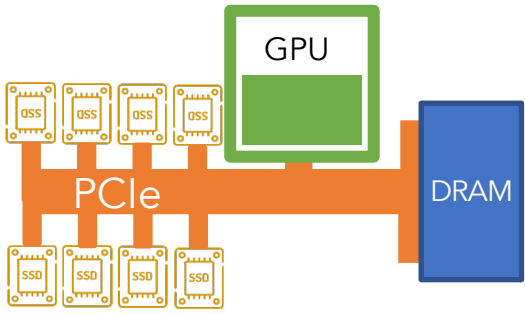


Topology provides options && workload guides decisions

Impact of Indirect NVMe-to-GPU Transfers

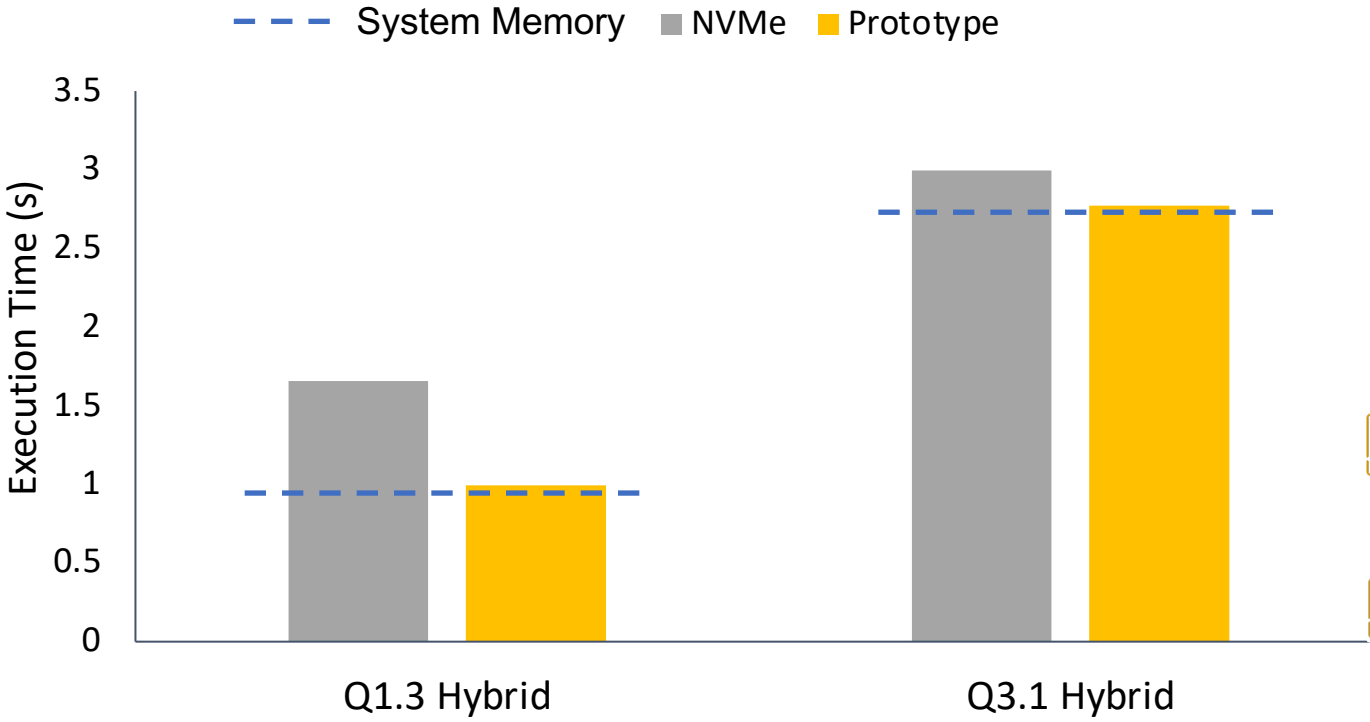
24-core AMD EPYC 7413
NVIDIA A40, PCIe 4.0 x16
8x (NVMe, PCIe 4 x4)
SSB sf=1000



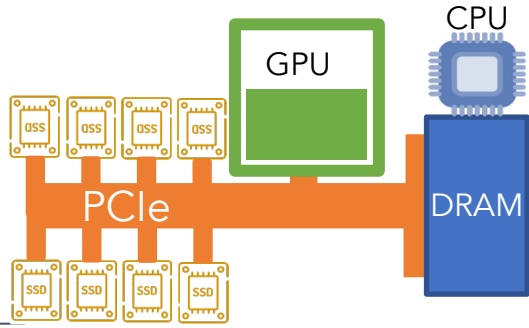


Indirect NVMe->GPU transfers improve execution by 13-45%

Bringing it All Together



24-core AMD EPYC 7413
NVIDIA A40, PCIe 4.0 x16
8x (NVMe, PCIe 4 x4)
96GB working set
10GB max GPU memory
90GB max CPU memory



25%-100% less DRAM required

Conclusion

- Storage bandwidth is approaching memory bandwidth
  - In-memory performance without (all of) the memory
- GPUs break the hierarchy
  - Yet, opportunities to improve resource utilization
- Variable hardware topology and workload
  - Specialize and adapt at runtime

Thank You!

www.nicholson.ai
hamish.nicholson@epfl.ch