AQUAQ Analytics

Optane

Experts in fast data solutions

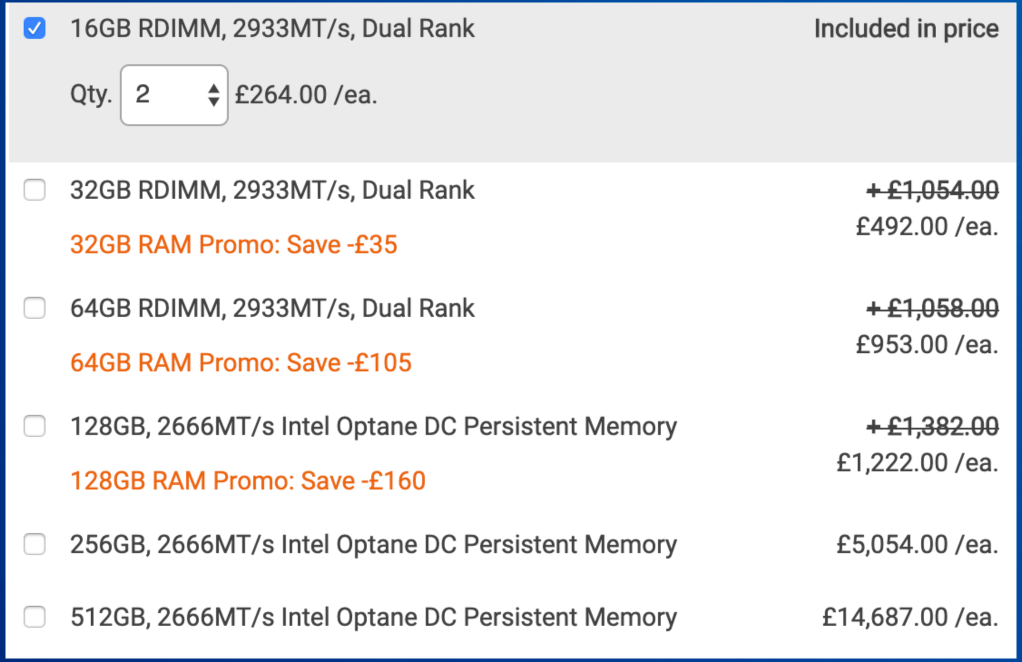
for demanding environments

AQUAQ
Analytics

Optane AppDirect support

What is Optane?

- New type of RAM –NVRAM (Non-Volatile) or persistent RAM
- Unlike DRAM, contents persist after machine restart – more like disk
- Slower than DRAM, much faster than SSD
- Cheaper than DRAM and maximum size per chip much higher



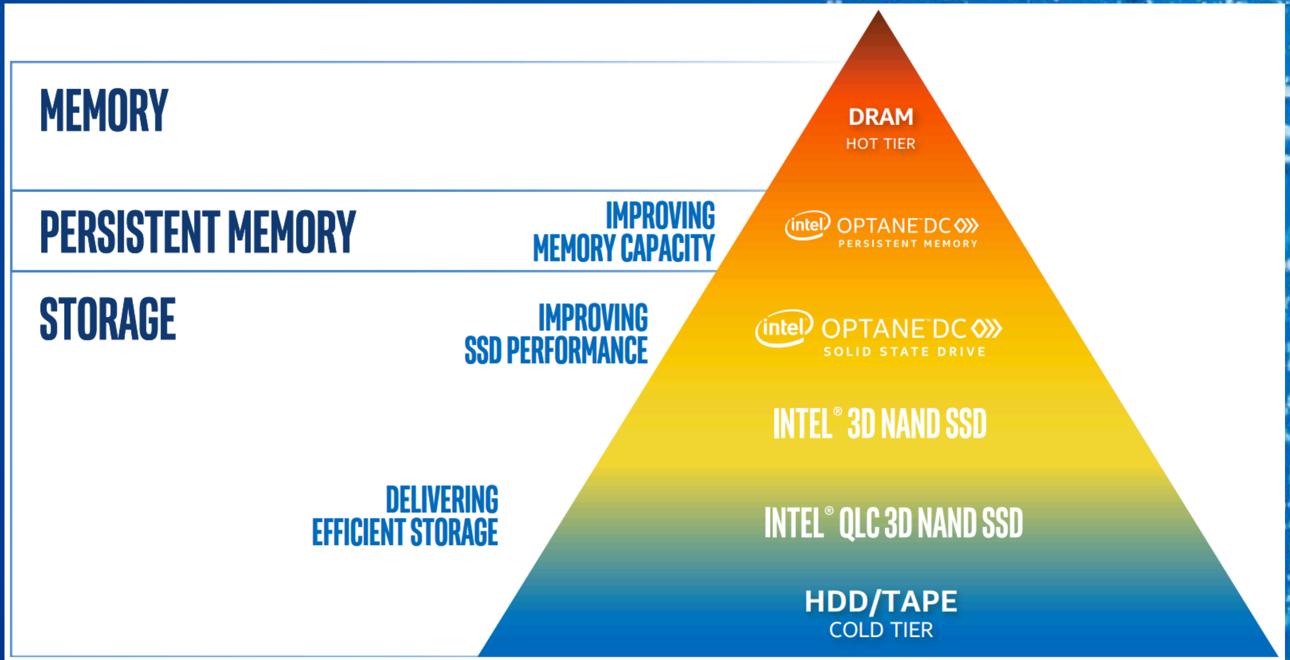
☑ 16GB RDIMM, 2933MT/s, Dual Rank Qty. 2 £264.00 /ea.	Included in price
☐ 32GB RDIMM, 2933MT/s, Dual Rank 32GB RAM Promo: Save -£35	~~+ £1,054.00~~ £492.00 /ea.
☐ 64GB RDIMM, 2933MT/s, Dual Rank 64GB RAM Promo: Save -£105	~~+ £1,058.00~~ £953.00 /ea.
☐ 128GB, 2666MT/s Intel Optane DC Persistent Memory 128GB RAM Promo: Save -£160	~~+ £1,382.00~~ £1,222.00 /ea.
☐ 256GB, 2666MT/s Intel Optane DC Persistent Memory	£5,054.00 /ea.
☐ 512GB, 2666MT/s Intel Optane DC Persistent Memory	£14,687.00 /ea.

AQUAQ
Analytics

Optane AppDirect support

Optane Performance

DDR4 memory accesses – 14ns
Optane DIMM – 350ns
NVMe Optane SSD access can take 10,000ns (10µs)
NVMe NAND SSD write – 30,000ns (30µs)
NVMe NAND SSD read – 120,000ns (120µs)
SATA NAND SSD read – 500,000ns (500µs or 0.5ms)
SATA NAND SSD write – 3,000,000ns (3,000µs or 3ms)
Disk drive seek – 100,000,000ns (100,000µs or 100ms)



MEMORY
PERSISTENT MEMORY
STORAGE
IMPROVING MEMORY CAPACITY
IMPROVING SSD PERFORMANCE
DELIVERING EFFICIENT STORAGE
DRAM
HOT TIER
intel OPTANE DC PERSISTENT MEMORY
intel OPTANE DC SOLID STATE DRIVE
INTEL® 3D NAND SSD
INTEL® QLC 3D NAND SSD
HDD/TAPE
COLD TIER

Optane AppDirect support

Optane Modes

- Storage mode – Optane presents as a disk
- Memory mode – Optane becomes main memory pool, DRAM is L4 cache
- AppDirect mode – DRAM and Optane present as separate memory pools to applications

Optane AppDirect support

Storage Mode

- Applications talk to Optane through a file system interface
- Optane just behaves as very fast SSD
- File API impacts performance – not as fast as direct access
- Transparent to kdb+, compatible with all versions

Optane AppDirect support

Memory Mode

- Optane becomes the main memory pool – so a machine with 256Gb of DRAM and 1Tb of Optane would appear to applications as having 1Tb RAM
- DRAM is managed as cache transparently by OS
- Frequently accessed objects in memory should remain in DRAM with occasional cache misses which have to hit Optane
- Compatible with all versions of kdb+ - could potentially have huge 1Tb+ RDBs if caching performance is suited to how kdb+ is accessing in memory data

AQUAQ
Analytics

Optane AppDirect support

App Direct Mode

- Applications can see DRAM and Optane pools separately
- Software must be rewritten to take advantage of Optane memory
- Kdb+ 4.0 adds support for App Direct mode with compatible Intel processor (Cascade Lake - 2020)

AQUAQ
Analytics

Benchmarks - setup

- Intel loaner machine – 48 core, 384Gb DRAM, 6Tb Optane
- Generated quotes and trades database
- 4096 syms (`aaa, `aab, … , `ppp)
- 1.3B quotes, 300M trades per day, 5 days
- On disk: 70G per day, 350G total

AQUAQ Analytics

Benchmarks - setup

```
[user1@atsnode24 ~]$ sudo ipmctl show -topology
[sudo] password for user1:
 DimmID | MemoryType                  | Capacity    | PhysicalID| DeviceLocator
================================================================================
 0x0001 | Logical Non-Volatile Device | 502.563 GiB | 0x0026    | CPU1_DIMM_A2
 0x0011 | Logical Non-Volatile Device | 502.563 GiB | 0x0028    | CPU1_DIMM_B2
 0x0021 | Logical Non-Volatile Device | 502.563 GiB | 0x002a    | CPU1_DIMM_C2
 0x0101 | Logical Non-Volatile Device | 502.563 GiB | 0x002c    | CPU1_DIMM_D2
 0x0111 | Logical Non-Volatile Device | 502.563 GiB | 0x002e    | CPU1_DIMM_E2
 0x0121 | Logical Non-Volatile Device | 502.563 GiB | 0x0030    | CPU1_DIMM_F2
 0x1001 | Logical Non-Volatile Device | 502.563 GiB | 0x0032    | CPU2_DIMM_A2
 0x1011 | Logical Non-Volatile Device | 502.563 GiB | 0x0034    | CPU2_DIMM_B2
 0x1021 | Logical Non-Volatile Device | 502.563 GiB | 0x0036    | CPU2_DIMM_C2
 0x1101 | Logical Non-Volatile Device | 502.563 GiB | 0x0038    | CPU2_DIMM_D2
 0x1111 | Logical Non-Volatile Device | 502.563 GiB | 0x003a    | CPU2_DIMM_E2
 0x1121 | Logical Non-Volatile Device | 502.563 GiB | 0x003c    | CPU2_DIMM_F2
 N/A    | DDR4                        | 32.000 GiB  | 0x0025    | CPU1_DIMM_A1
 N/A    | DDR4                        | 32.000 GiB  | 0x0027    | CPU1_DIMM_B1
 N/A    | DDR4                        | 32.000 GiB  | 0x0029    | CPU1_DIMM_C1
 N/A    | DDR4                        | 32.000 GiB  | 0x002b    | CPU1_DIMM_D1
 N/A    | DDR4                        | 32.000 GiB  | 0x002d    | CPU1_DIMM_E1
 N/A    | DDR4                        | 32.000 GiB  | 0x002f    | CPU1_DIMM_F1
 N/A    | DDR4                        | 32.000 GiB  | 0x0031    | CPU2_DIMM_A1
 N/A    | DDR4                        | 32.000 GiB  | 0x0033    | CPU2_DIMM_B1
 N/A    | DDR4                        | 32.000 GiB  | 0x0035    | CPU2_DIMM_C1
 N/A    | DDR4                        | 32.000 GiB  | 0x0037    | CPU2_DIMM_D1
 N/A    | DDR4                        | 32.000 GiB  | 0x0039    | CPU2_DIMM_E1
 N/A    | DDR4                        | 32.000 GiB  | 0x003b    | CPU2_DIMM_F1
[user1@atsnode24 ~]$
[user1@atsnode24 ~]$
[user1@atsnode24 ~]$ sudo ipmctl show -memoryresources
 MemoryType   | DDR         | PMemModule   | Total
==========================================================
 Volatile     | 381.500 GiB | 0.000 GiB    | 381.500 GiB
 AppDirect    | -           | 6024.000 GiB | 6024.000 GiB
 Cache        | 0.000 GiB   | -            | 0.000 GiB
 Inaccessible | 2.500 GiB   | 7.184 GiB    | 9.684 GiB
 Physical     | 384.000 GiB | 6031.184 GiB | 6415.184 GiB
```

Benchmarks - setup

```
ipmctl create -goal persistentmemorytype=appdirect

ndctl create-namespace --mode=fsdax --region=0 --size=2052G --align=2M

mkfs -t xfs /dev/pmem0

mkdir /mnt/pmem
mount -o dax /dev/pmem0 /mnt/pmem/
chmod 777 /mnt/pmem
```

Benchmarks – Optane vs DRAM

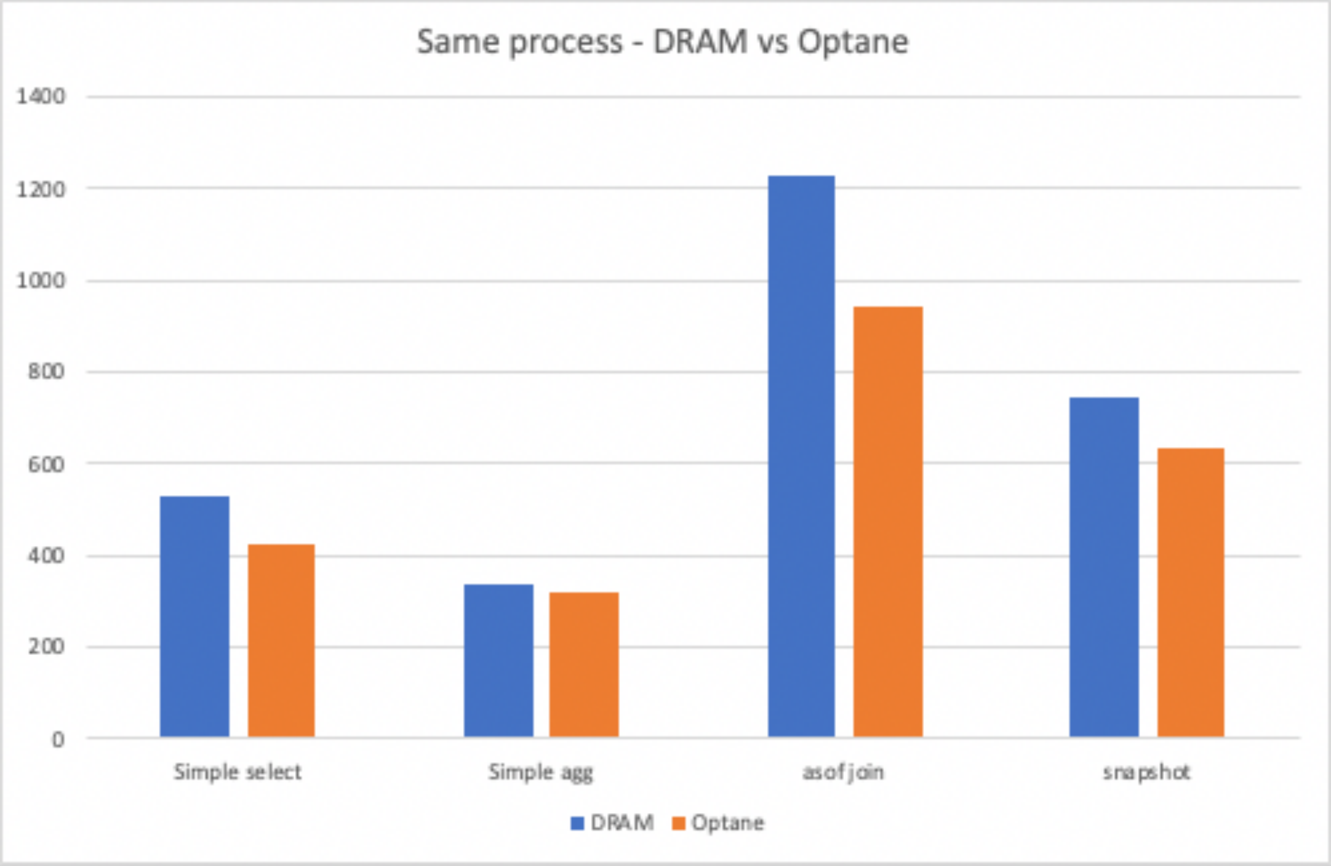
```
/ Queries

/ simple select
\t:100 select from trades where sym in -5?`3
/ simple agg
\t:100 select avg price, sum size by sym from trades where sym in -10?`3
/ asof join
\t:10 aj[`sym`time;select from trades where sym in -5?`3;quotes]
/ snapshot
\t:10 select by sym from trades where sym in -100?`3,time<=2014.04.21D10
```

Benchmarks – Optane vs DRAM

Scenario 1

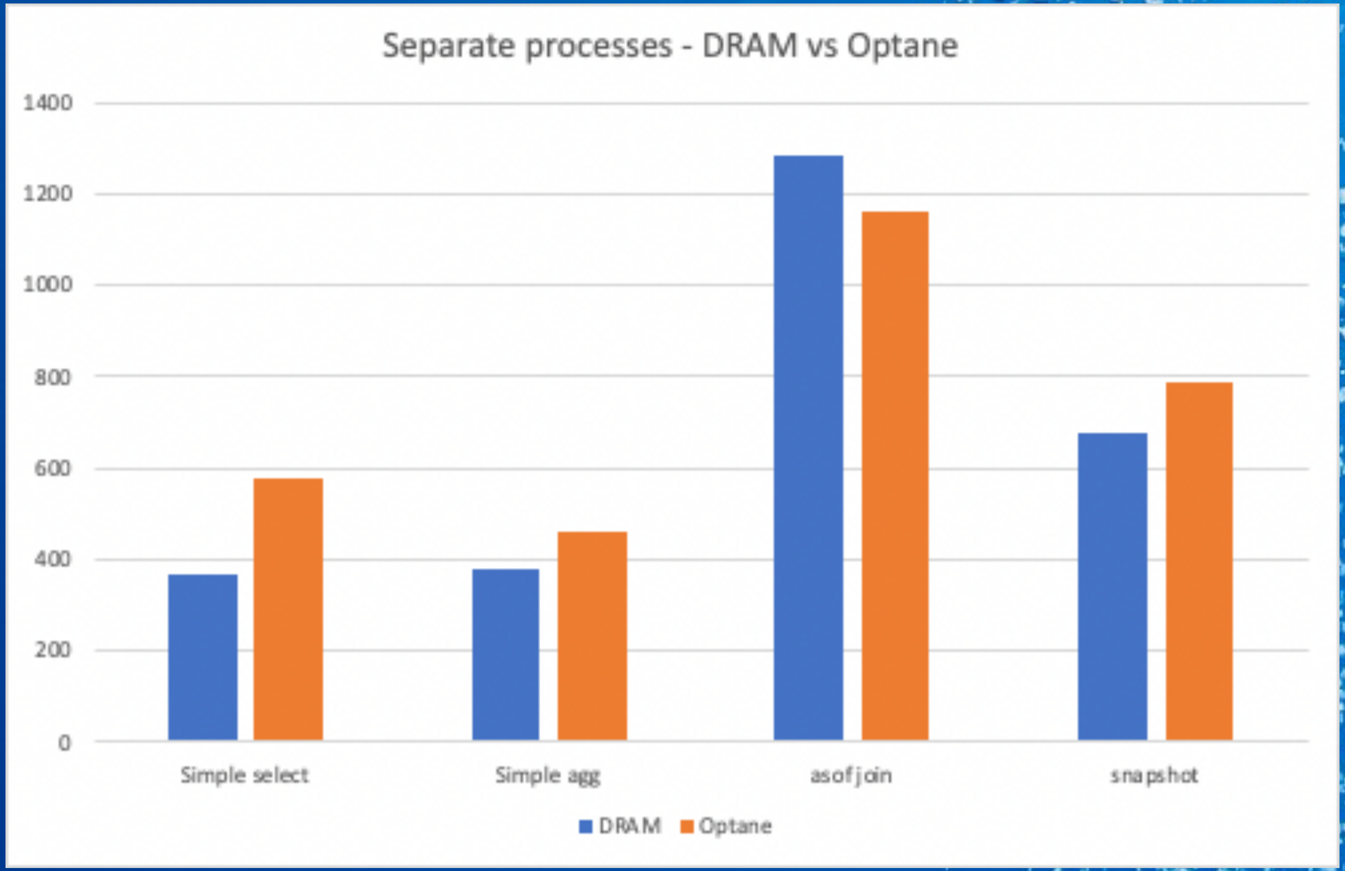
- Single kdb+ process:
  - 1 day of data in DRAM
  - 1 day of data in Optane



Benchmarks – Optane vs DRAM

Scenario 2

- Separate kdb+ processes:
  - 1 day of data in DRAM
  - 1 day of data in Optane



Benchmarks – Optane vs Disk

- 5 days history loaded in Optane
- Each table is nested by date:

```
q)trades
2014.04.21| +`sym`time`src`price`size!(`p#`aaa`aaa`aaa`aaa`aaa`aaa`aaa`aaa`aa..
2014.04.22| +`sym`time`src`price`size!(`p#`aaa`aaa`aaa`aaa`aaa`aaa`aaa`aaa`aa..
2014.04.23| +`sym`time`src`price`size!(`p#`aaa`aaa`aaa`aaa`aaa`aaa`aaa`aaa`aa..
2014.04.24| +`sym`time`src`price`size!(`p#`aaa`aaa`aaa`aaa`aaa`aaa`aaa`aaa`aa..
2014.04.25| +`sym`time`src`price`size!(`p#`aaa`aaa`aaa`aaa`aaa`aaa`aaa`aaa`aa..
q){d:.Q.w[];system"d .m";o:.Q.w[];system"d .";(`used`mphy#(d;o))%1024 xexp 3}[]
used         mphy
---------------------
0.0003374666 375.1356
552          2018.949
```

AQUAQ Analytics

Benchmarks – Optane vs Disk

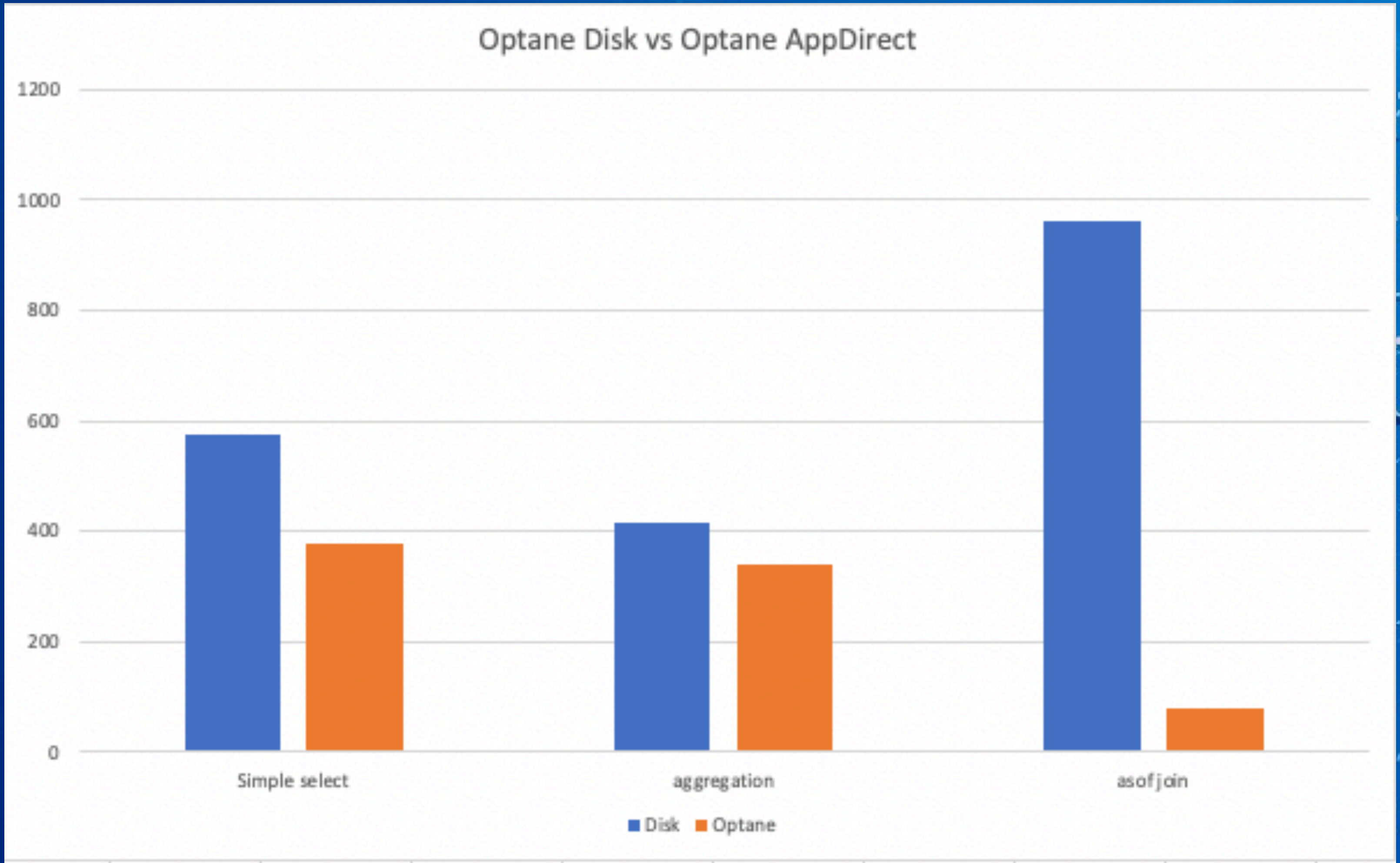
```
/ Queries

/ simple select - 3 days, 10 syms
\t select from quotes where date in -3?date, sym in -10?`3
\t raze {select from quotes[x] where sym in -10?`3} each -3?key quotes

/ aggregation - 3 days, 5 syms, 1h bars
\t select last bid, last ask by date, sym, 0D1 xbar time from quotes where date in -3?date, sym in -5?`3
\t raze {select last bid, last ask by sym, 0D1 xbar time from quotes[x] where sym in -5?`3} each -3?key quotes

/ asof lookup - 1 day, rack of 3000 sym & time samples
rack:([]time:2014.04.23D09:30+0D00:05*til 60)cross([]sym:-50?`3)
\t:50 aj[`sym`time;rack;select time,sym,bid,ask from quotes where date=2014.04.23]
\t:50 aj[`sym`time;rack;quotes[2014.04.23]]
```

Benchmarks – Optane vs Disk



AQUAQ Analytics

Conclusions

- Performance is great – Optane Appdirect delivers DRAM-like performance for most kdb+ workloads tested
- Largest Optane chips are expensive, but deliver memory in a volume not possible with DRAM
- Flexible – can be chopped up into partitions to separate process memory pools and used as disk at the same time
- Kdb+ architectures will have to change to take advantage of Optane

AQUAQ
Analytics

Q&A