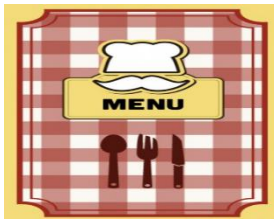
•Building an Open Memory-Centric Computing Architecture using Intel Optane

•Frank Ober
•Efstathios Efstathiou

•Oracle Open World 2017 – October 3, 2017

Open MCCA

Agenda



- The legal stuff
- Why Memory Centric Computing?
- Overview of Open Memory Centric Computing Architecture (OpenMCCA)
- Optane SSDs performance capabilities (Frank Ober, Intel)
- OpenMCCA: Technical Demos
- Summary with Q/A

The legal stuff

Disclaimer for Efstathios Efstathiou

- In this presentation I express my view of things based on my expertise as gathered as Expert for Oracle Technology, which may in some aspects be different from the strategic decisions of my employer
- Part of my research is based on project work for the Swiss Government, while some work and studies were done in my personal free time for my personal education
- OpenMCCA was developed entirely on my personal free time

Open MCCA

Intel Legal Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit .

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo, Xeon, Intel® Optane™, and 3D XPoint are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

•Why Memory Centric Computing?

•History

- Started as Engineering task to better understand Infiniband
- Create an NVMe Storage Server using a PCIe Switch Chip
- Experimented with PCIe NTB and Device Sharing
- Goal is to build Fabric Attached Memory Concept

•Engineering Work (Work Time)

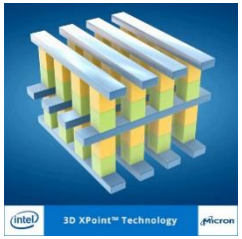
•Saturday's Club Work (Private Time)

INFINIBAND TRADE ASSOCIATION

nvm EXPRESS

flashgrid

PLEXISTOR



PCI EXPRESS

•Open
•MCC
A

•2016

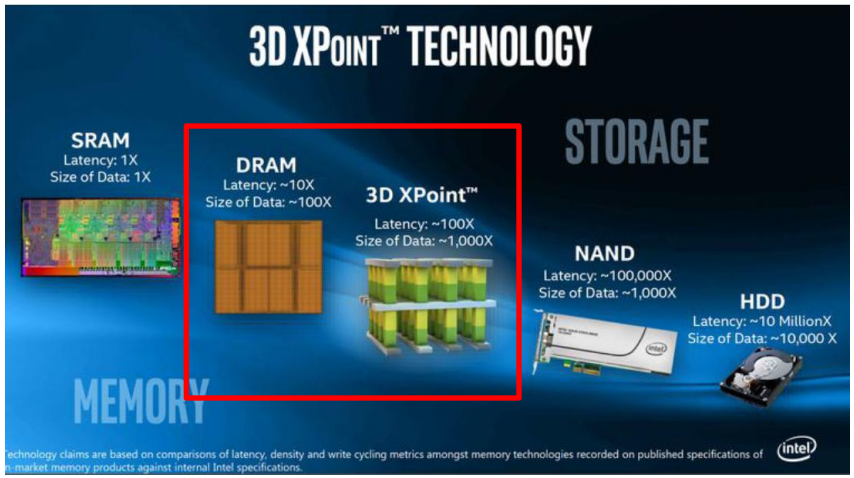
2017

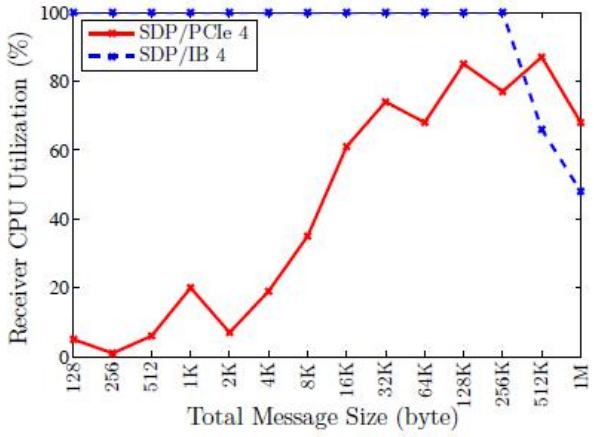
Open MCCA

•Why Memory Centric Computing?

•Technical toughts

- Ideally everything would be memory
- Network ist the bottleneck (latency)
- Multiple protocols (transport, storage, interconnect)
- Why not use what's already there?

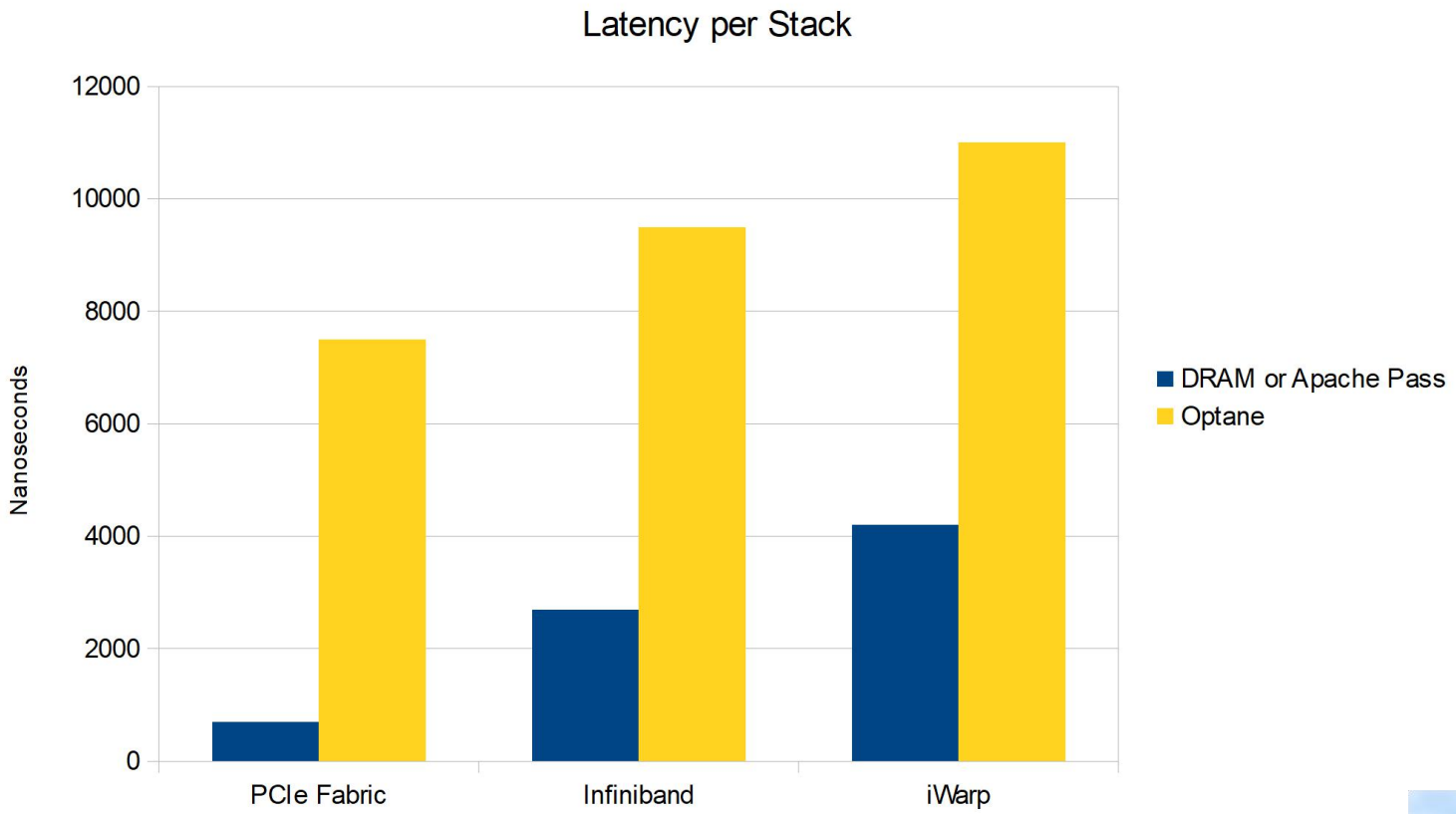




(g) Receiver, four concurrent connections

•Why Memory Centric Computing?

•Technical toughts

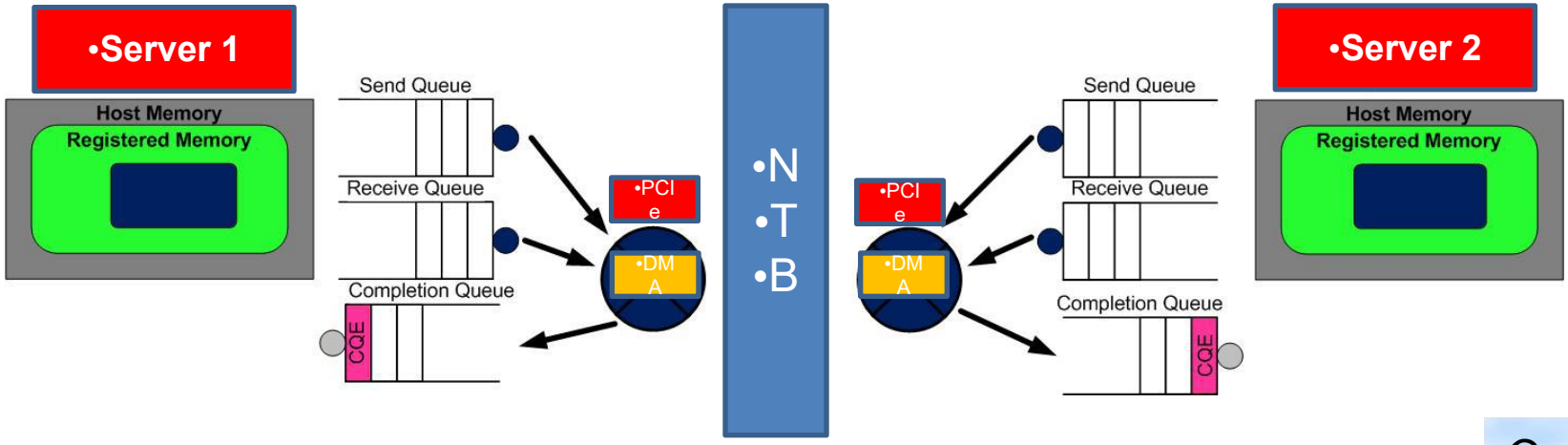


•Open
•MCC

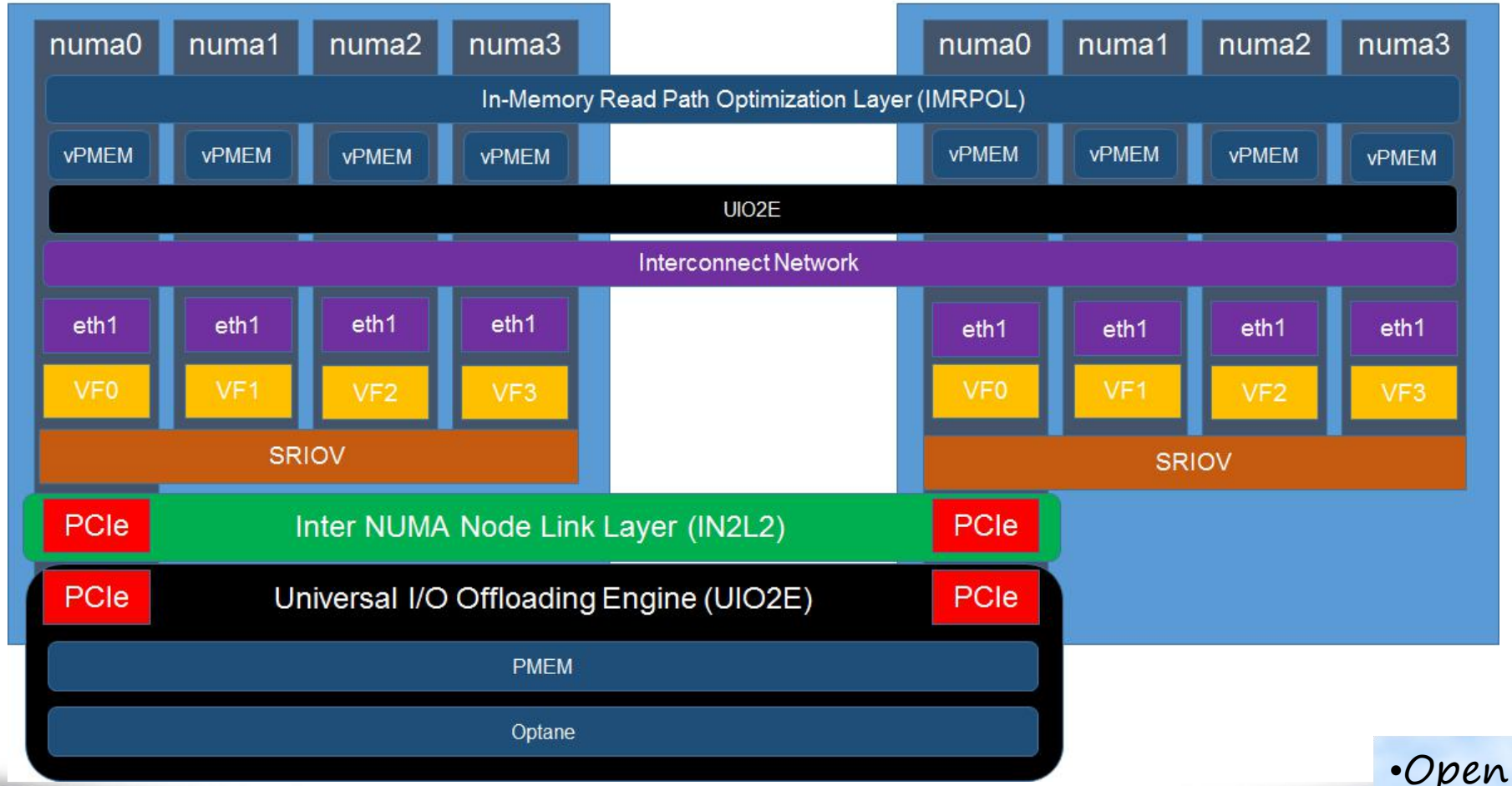
•Why Memory Centric Computing?

•Decisions

- Use PCIe as a unified fabric
- Use DMA engines to offload I/O work like we have with IB
- Combine tiering and remote adressable memory



•Overview of Open Memory Centric Computing Architecture (OpenMCCA)



•Open
•MCC

•Overview of Open Memory Centric Computing Architecture (OpenMCCA)

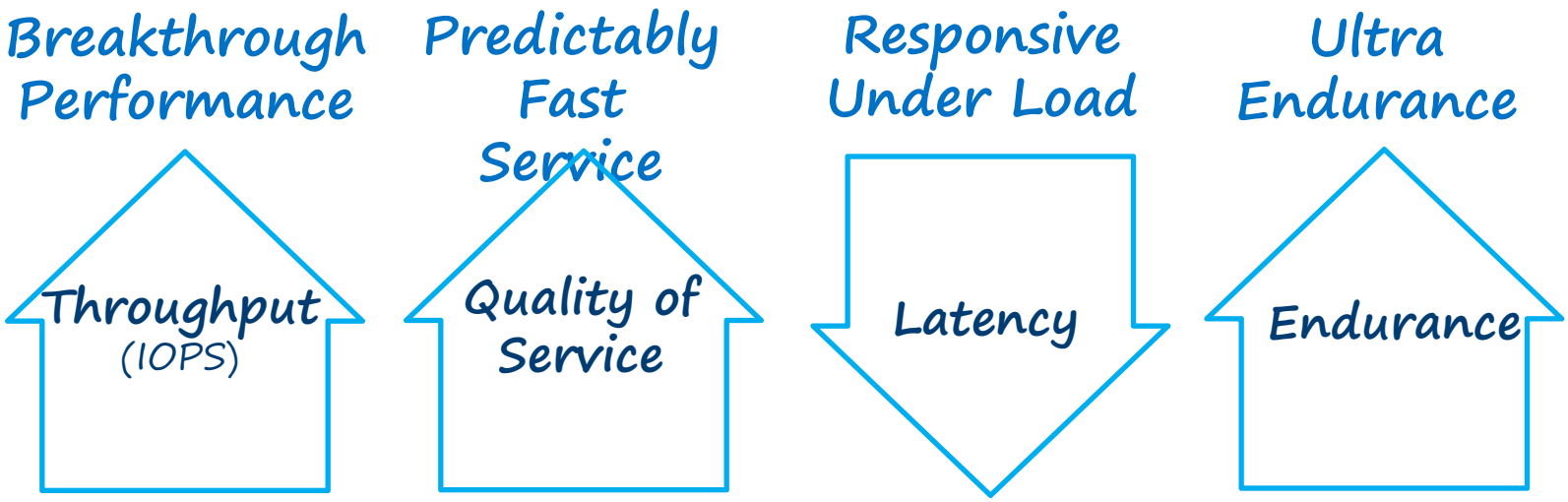
Features

- In-Memory Read Path Optimization (IMRPOL)
  - Access the fastest copy of your data using the shortest path in the best tier when using Software Defined Memory
- Universal I/O Offloading Engine
  - Save CPU cycles and licenses on your database host
  - Optimize your data center rack design
  - Add hardware accelerators on the fly using PCIe Device lending / MR-IOV
  - OptaneGRID I/O-modules for maximum performance

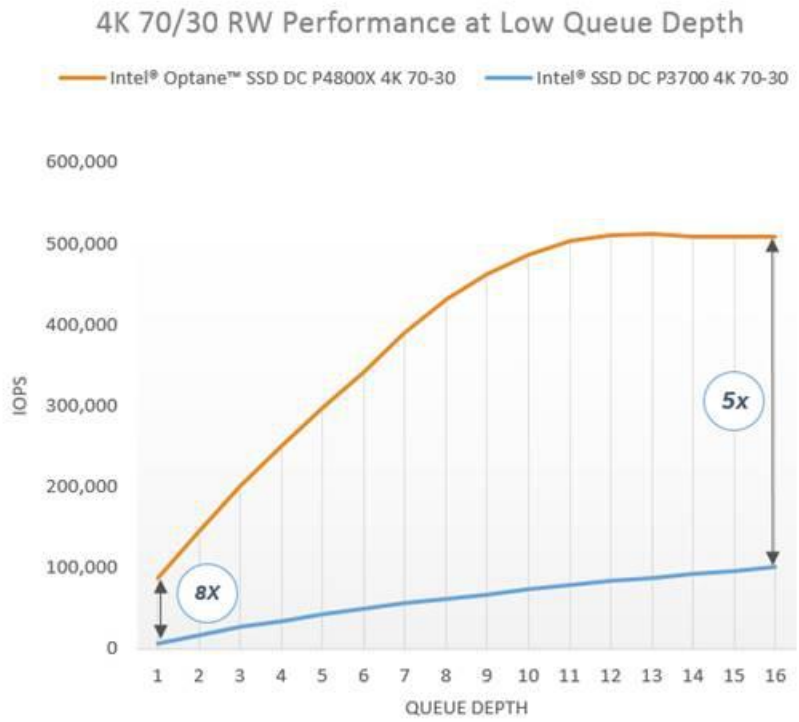
Open MCCA

Optane SSDs performance capabilities

Intel® Optane™ SSD DC P4800X



•Breakthrough Performance

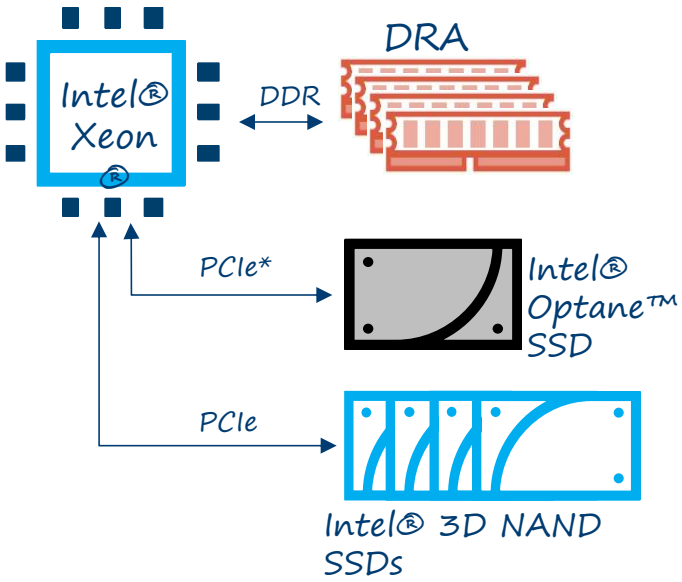


- •5-8x faster at low Queue Depths[1]
- •Vast majority of applications generate low QD storage workloads

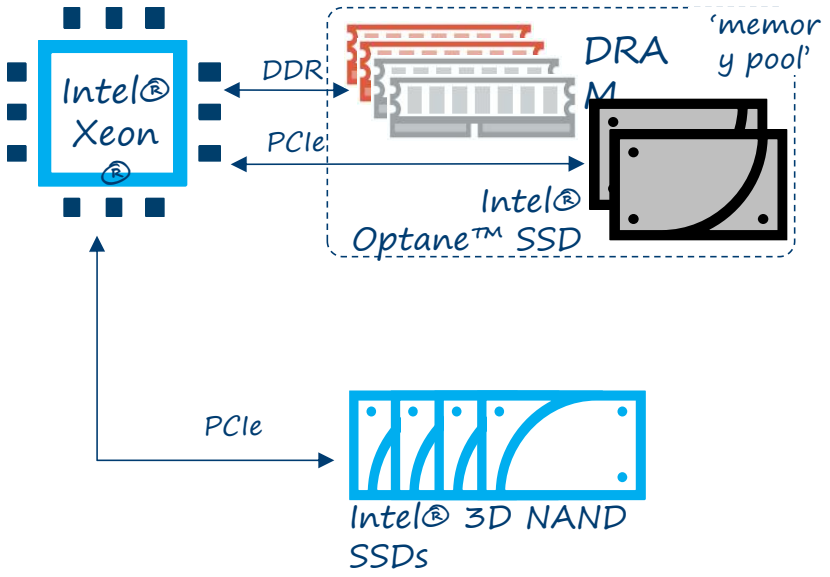
1. Common Configuration - Intel 2U Server System, OS CentOS 7.2, kernel 3.10.0-327.el7.x86_64, CPU 2 x Intel® Xeon® E5-2699 v4 @ 2.20GHz (22 cores), RAM 396GB DDR @ 2133MHz. Configuration – Intel® Optane™ SSD DC P4800X 375GB and Intel® SSD DC P3700 1600GB. Performance – measured under 4K 70-30 workload at QD1-16 using fio-2.15. Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance.

Intel® Optane™ SSD Use Cases

Fast Storage



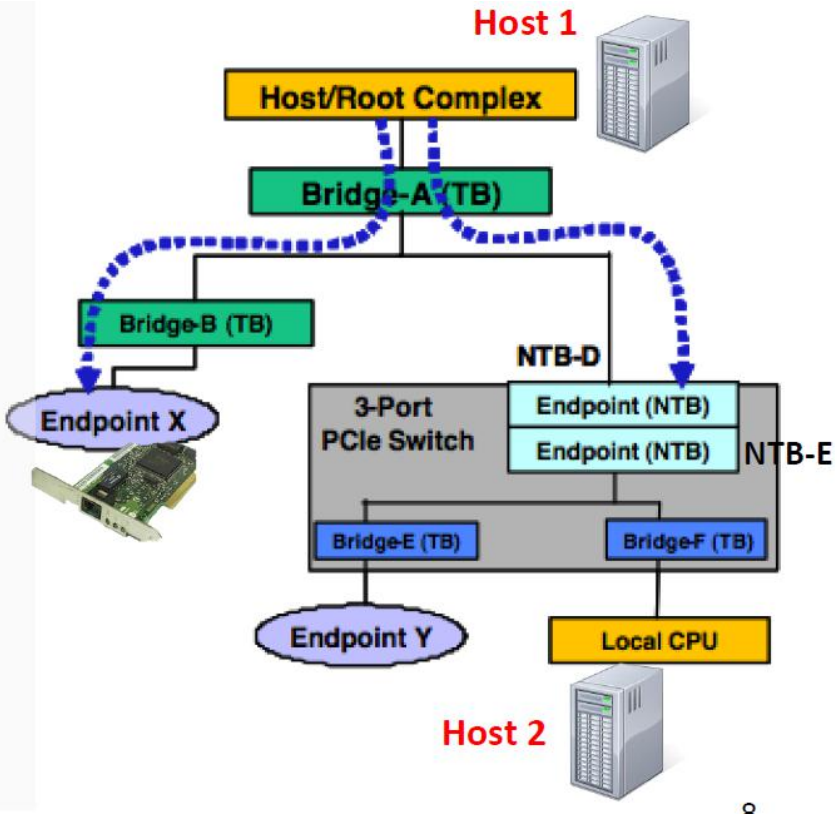
Extend Memory



*Other names and brands names may be claimed as the property of others

•OpenMCCA: Technical Demos

Setup for Demo 1: Diagram



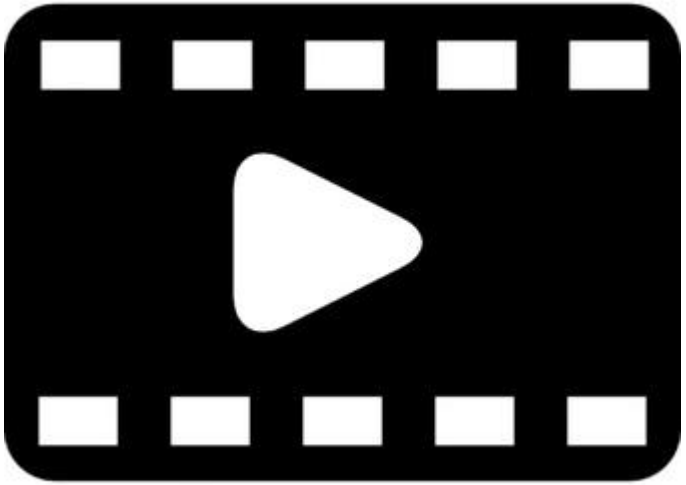
•OpenMCCA: Technical Demos

Setup for Demo 1: Technical Explanation

- PCIe Device Sharing
  - Node 1 has an Intel quad port PCIe NIC
  - Node 1 is acting as device lending host (target)
  - Node 1 enabled SRIOV on the nic and allows sharing through PCIe
  - Node 2 acts as client and can now use the nic from Host1 trough PCIe

•OpenMCCA: Technical Demos

•Demo 1: PCIe Device Sharing



•Our inspiration

•https://www.youtube.com/watch?v=GPh0Ms3dfPo

Open MCCA

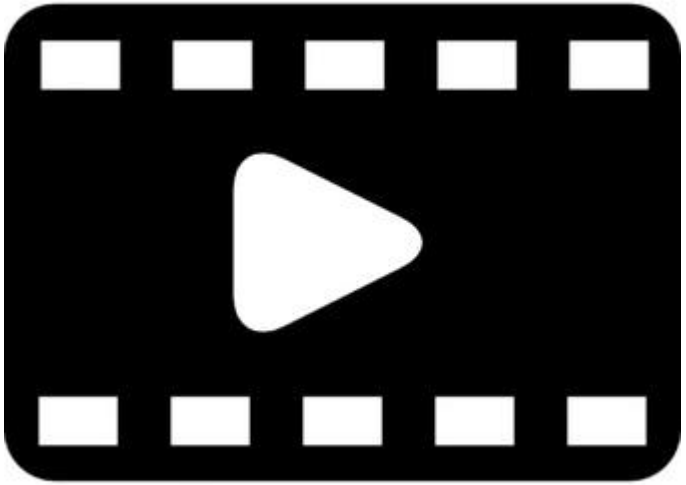
•OpenMCCA: Technical Demos

Setup for Demo 2: Technical Explanation

- FIO Benchmark
  - Node 1 has a direct attached PCIe switch based NVMe JBOF containing a single OptaneGrid Module
  - Node 1 is the client executing fio

•OpenMCCA: Technical Demos

•Demo 2: Local Storage Tier Performance



•Single Node Performance, with «Wasted Cores» Scheduler

•https://github.com/Turbine1991/build_ubuntu_kernel_wastedcores

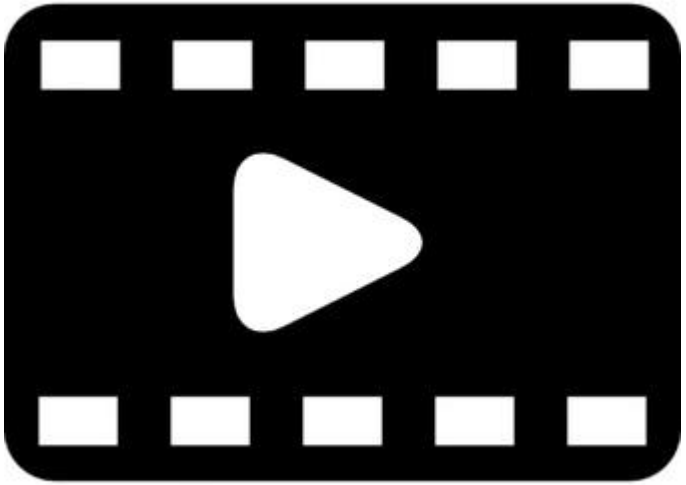
•OpenMCCA: Technical Demos

Setup for Demo 3: Technical Explanation

- FIO Benchmark
  - Node 1 has a direct attached PCIe switch based NVMe JBOF containing a single OptaneGrid Module
  - Node 1 is the device lending host (target)
  - Node 2 is accessing Host 1's locally attached NVMe devices trough PCIe NTB
  - Node 2 is the client executing fio

•OpenMCCA: Technical Demos

•Demo 3: I/O Offloading



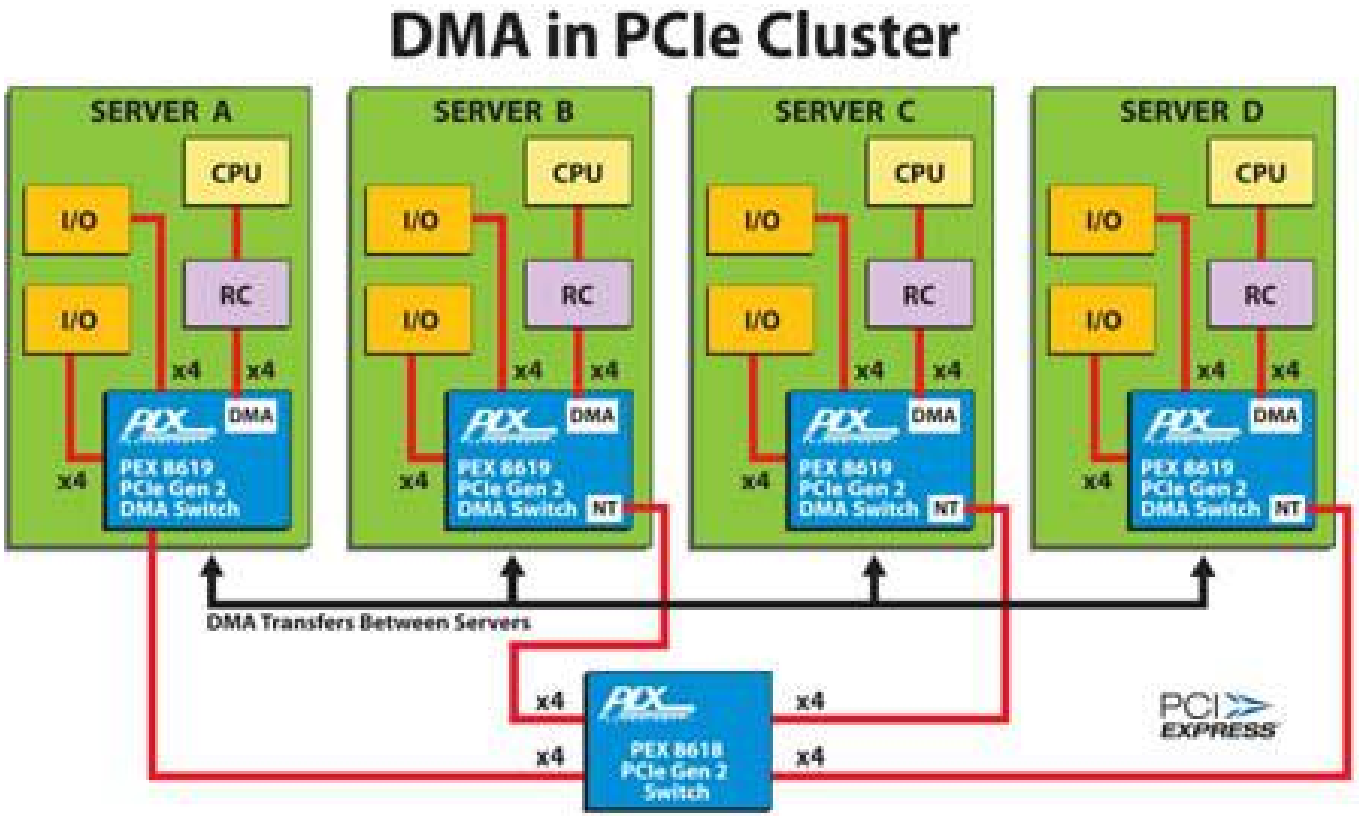
•Use PCI Switch DMA Engine

•https://stackoverflow.com/questions/27470885/how-does-dma-work-with-pci-express-devices

Open MCCA

•OpenMCCA: Technical Demos

•Setup for Demo 4: Diagram



Open MCCA

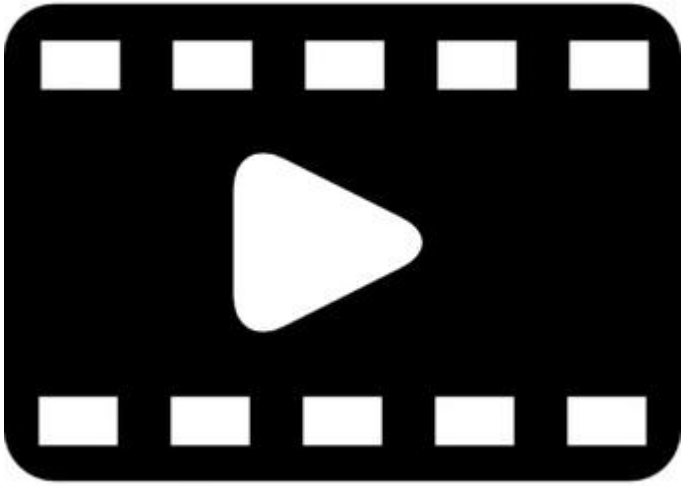
•OpenMCCA: Technical Demos

Setup for Demo 4: Technical Explanation

- Oracle RAC Cluster with ASM
  - 2 DB nodes with 8 hard partitioned cpu cores each
    - 1x PCIe Fabric Switch
    - 4x Device/Memory lending host
    - Using OptaneGrid devices for Memory Expansion
    - DB-Server DRAM as Cache mirrored using PCIe NTB

•OpenMCCA: Technical Demos

•Demo 4: Oracle RAC + OpenMCAA



•Aggregate multiple nodes using Oracle RAC and ASM

Open MCCA

•OpenMCCA: Technical Demos

•Demo 4: Oracle RAC + OpenMCAA

•RESULTS

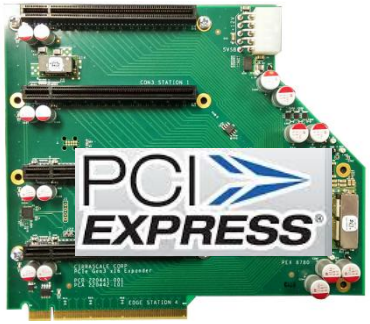
```
[oracle@vscale07 ~]$ f_bench_rac

SQL*Plus: Release 12.2.0.1.0 Production on Mon Sep 25 22:17:43 2017

Copyright (c) 1982, 2016, Oracle.  All rights reserved.


Verbunden mit:
Oracle Database 12c Enterprise Edition Release 12.2.0.1.0 - 64bit Production

SQL>
max_iops = 5239220
latency  = 0
max_mbps = 202623

PL/SQL-Prozedur erfolgreich abgeschlossen.

SQL> Verbindung zu Oracle Database 12c Enterprise Edition Release 12.2.0.1.0 - 64bit Production beendet
```

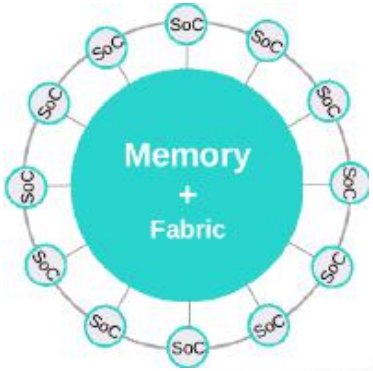
•Aggregate multiple nodes using Oracle RAC and ASM

Open MCCA

•Summary

- OpenMCCA Fabric Attached Memory using Intel Optane give you:
  - **Best in class per licensed cpu core performane (25 GB/s)**
  - Using:
    - **Less Hardware**
    - **Less SGA Memory (data is already in DRAM)**
- **PCIe is the „Mother of all Fabrics“** (unbeaten in latency)
- OptaneGRID 3DXpoint can further cut latency getting very close to DRAM (700ns)





•Open
•MCCA

Open
MCCA

•Q&A


IDEA KILLER
BINGO
But...
We've already tried that before.
It'll never fly.
Let me play devil's advocate here...
Let's not go off on a tangent.
You're setting yourself up for failure.
Sure it will...
In THIS economy?
Do you think we're made of money?
That's not a high priority right now.
Have you really thought about the implications?
That won't work because...
FREE
The only problem with that is...
Run an ROI, and get back to us.
Is this in line with our strategy?
The front line will never go for it.
You're kidding... right?
Yes, but...
Does anyone really care about that?
What you are really saying...
If it ain't broke...
Sure, in theory... but you don't think it'll really work
But how much is this idea worth?
Do we really have the resources for this?

Open MCCA

•Thanks to our supporters

Open
MCCA

•Special thanks go to

Tim Reuter aka „Gyro Gearloose”

Lars Kristiansen

Alain Fournial

Amit Golander

Artem Danielov

Andy Du

Joerg Roskowetz

YT Huang

Stephen Bates

Gregory Eckert

Pilar Aguado

Gert Pauwels

Anastasios Panagou

Open MCCA


thank you
danke
謝謝
ngiyabonga
شكراً جزيلاً
рахмат
Баярлалаа
спасибо
teşekkür ederim
gracias
dank je
merci
obrigado
dziękuję
bedankt
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
grazie
arigatô
dakujem
terima kasih
ありがとう
감사합니다
ευχαριστώ
мерси
takk
taiku
sagolun
mauruuru
hvala
tapadh leat
mersi
barka
welalin
tack
spas
faafetai lava
vinaka
спасибі
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
дякую
djiere dieuf
tau
kiitos
dankie
dhanyavad
nanni
nandri
köszönöm
akun dankon
gràcie
bayarlalaa
enkosi
sobodi
dēkuji
mèsi
didi madloba
najis tuke
kam sah hamnida
rahmat
chnorakaloutioun
gratias ago
gràcies
sulpáy
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
trugarez
xièxie