FPGA Attached Persistent Memory

Accelerating Applications Cost-effectively through Coherency to meet Fast & *Predictable* Data needs

Sean In

Intel Corporation

Legal Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Intel, the Intel logo, Optane, Xeon, and Agilex are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

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

Introduction: Towards the Fast & *Predictable* Data

Then: Traditional and Big Data

- NAND Flash, SSD Controller, SSD, AFA Storage Systems
- DRAM: DDRx, HBMx, LPDDRx, GDDRx

Plus…

Now: Fast Data towards *Predictable* Data

- Persistent Memory, FPGA (Acceleration /Computation), Coherent & non-Coherent attached Memory & Storage product solutions

Critical Inflection Point based on Unpredictability & Complexity

- Traditional Data, Big Data, and Fast & *Predictable* Data
- New emergence of applications, use cases, and workloads (Enterprise/Cloud WLs vs AIML WLs)
- Deterministic transactions
- Compute, Storage, vs Memory-centric
- Storage vs Memory Semantics

- Heterogeneous Architecture (CPU, GPU, FPGA, ASIC)
- Interconnects (PCIe, CXL, CCIX, OpenCAPI)
- Volatile Memory (DDR, HBM, GDDR, LPDDR, SRAM)
- NAND Flash/ Storage (3D NAND, SLC/MLC/TLC/QLC, Low Latency NAND
- Persistent Memory (Optane, PCM, MRAM, ReRAM, FRAM, etc)
- Form Factors: Components, SiP, DIMM, SSD, Embedded, EDSFF

Addressing the Needs of Fast and *Predictable* Data

- Ever Increasing new types of Workloads
- Classification/ Recognition -> Prediction/ Decision
- Small, Random, Read
- Memory semantics
- Acceleration of Workloads (FPGA)
- Memory Hierarchy
- New types of Memory (Persistent Memory)
- Cache Coherency (Coherent Interconnect) / Standalone PCIe Interconnect

Not just about individual ingredients but

Product as Solution to address Fast Data Needs

Benefits of FPGA; Addressing the needs for Complex Workloads

MARKETS DEMANDING CUSTOMIZATION

EDGE
Real-Time Actionable Intelligence

NETWORK
High-Bandwidth Aggregation and Processing

DATA CENTER
Managing, Organizing, and Processing the Explosion of Data

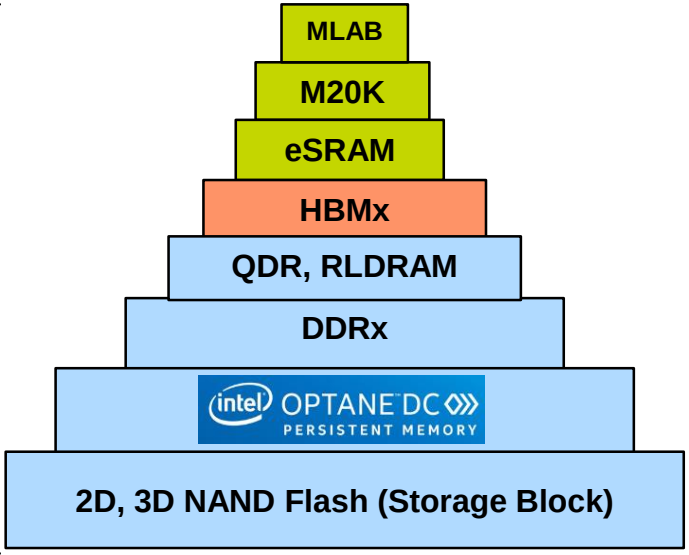
Flexibility
FPGA functionality can change upon every power-up of the device

Acceleration
TTM & Acceleration (Compression, Dedup, Classification, Encryption, etc)

Integration
on-die processors, transceiver I/O's, RAM blocks, DSP engines, and ***Different Memory hierarchies***

TCO
Benefits on NRE, Design cycles, Manufacturing cycles, etc

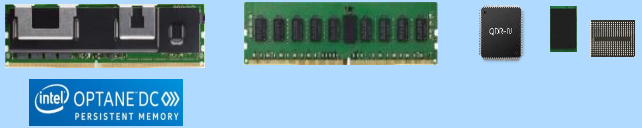
Memory Hierarchy for FPGA; Addressing Different Usage Needs



On-Chip Memory
Highest BW and Power Efficiency

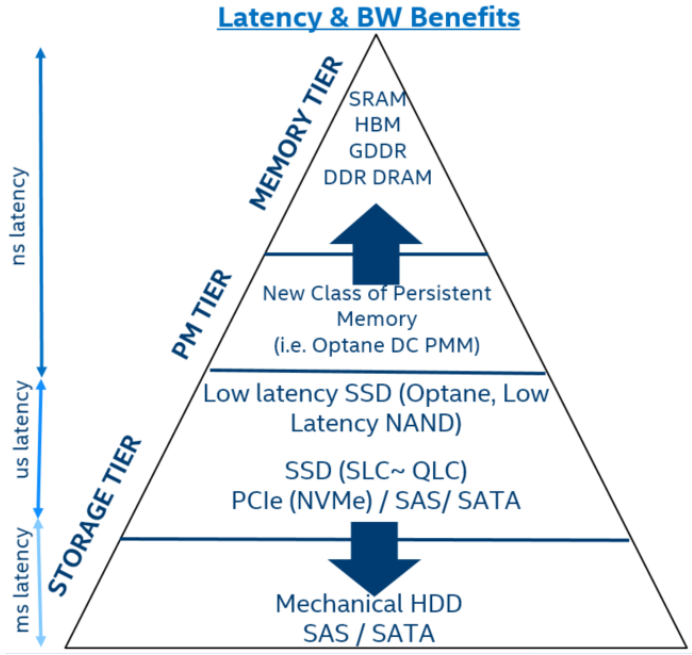
In-Package Memory
High Bandwidth, Density

On-Board
Flexibility, Density, Speed, Persistency



FPGA Memory & Storage Hierarchy Building Blocks

FPGA + Persistent Memory (PMEM) for Fast & *Predictable* Data needs

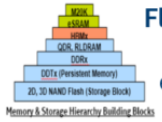


	FPGA+ PMEM	Values
Media	Persistent Memory (i.e. Optane)	Memory access with persistency
Memory I/0	DDR I/0	Higher I/O pin speed
I/F	PCIe, Coherent I/F (i.e. CXL or others)	Coherent attached
Capacity	FlexScale (small to large capacities)	Capacity Scaling
FF	AIC + DDR DIMM slot, On-board	FF Flexibility
Latency	Low	Transaction commit
BW	High	Performance

With FPGA Acceleration!

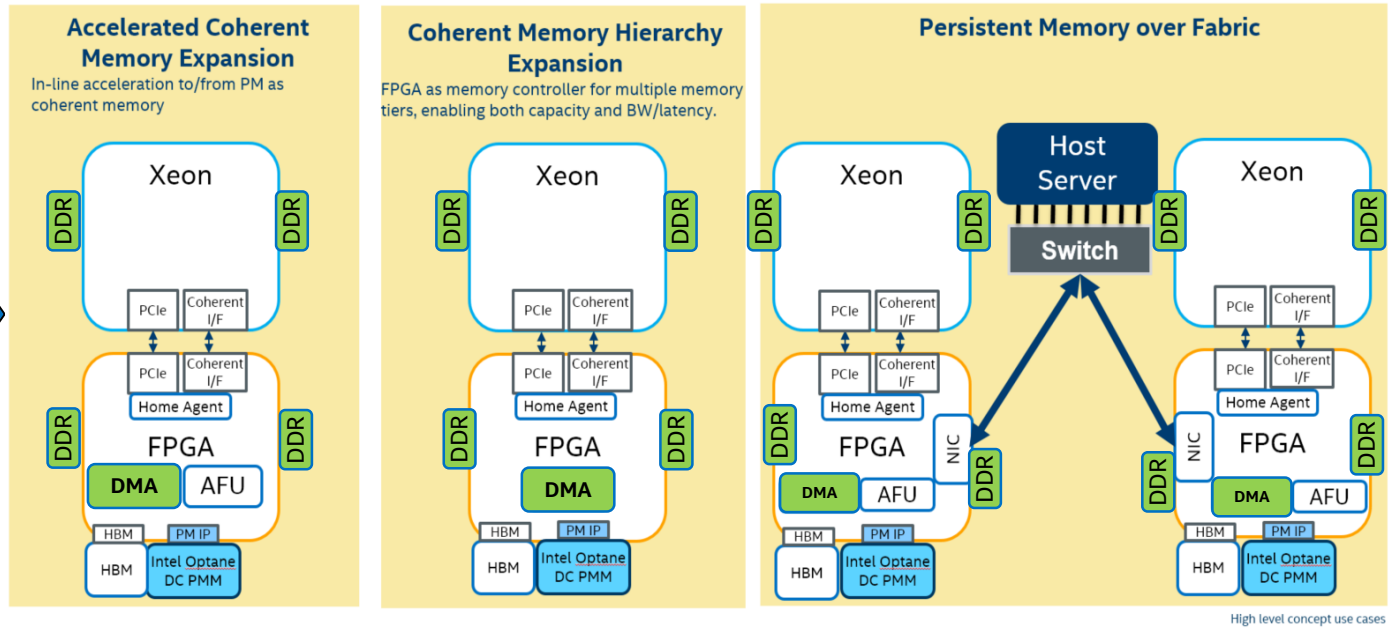
FPGA + PMEM Value Proposition

Addressing Fast and Big Data needs through Accelerated Coherent Memory Expansion

Coherent Memory Expansion	Accelerated (AFU) + Coherent Memory Expansion	Persistent Memory over Fabric
Larger Workloads 4TB+ capacity support per Device		
TCO Reduction Consolidation of different Memory and Storage tiers		
Performance for large data sets Memory transaction vs Block access/transaction		
 **Flexible Memory Usage Models** Enable Coherent memory expansion and leverage FPGA Memory Tiers	**Accelerate big data workloads** Enable in-line encryption, compression, analytics or other AFUs	**Scalable, Shared Memory** FPGA's Memory locality access over Fabric, scalable memory pool

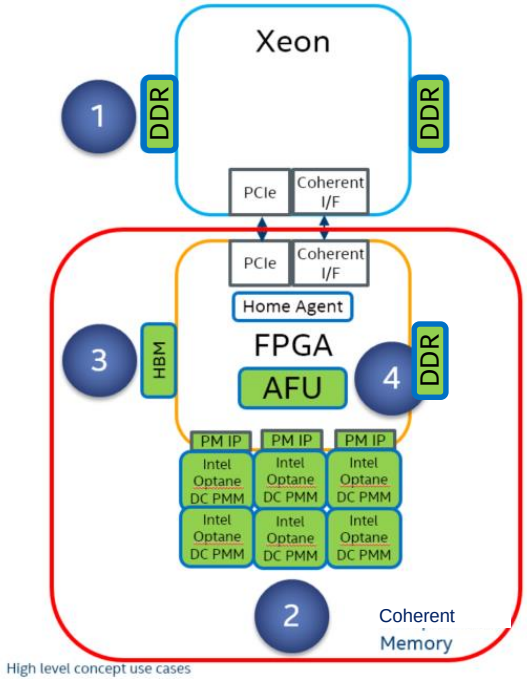
Coherent FPGA + PMEM Enables Variety of Use Cases

Different Coherent Topologies



Flash Memory Summit

Data Acceleration through minimal data movement & Memory semantics

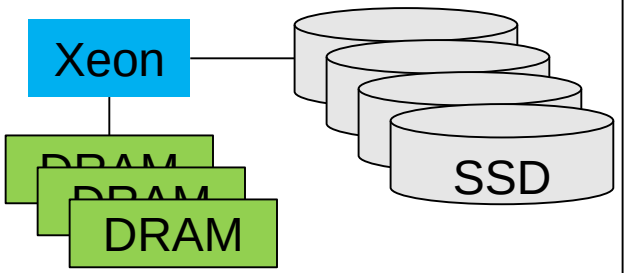


High level concept use cases

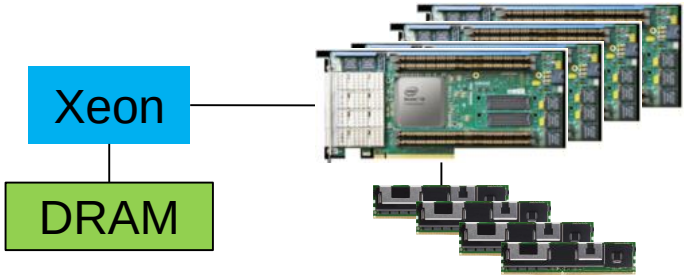
1. Minimize Data Movement from Storage to Memory through maximizing DRAM footprint - best performance for high SLA workloads
2. Dramatically increase over memory footprint and reduce Storage footprint
3. Smart caching Option to HBM for high bandwidth
4. Acceleration Functions – compression, encryption, query acceleration, others

intel

Intelligence Edge Metadata DB



- Big Meta DB in SSD
- Long DB loading time to DRAM
- IO BW busy



- Persistent Memory on FPGA to reduce DB loading time
- FPGA acceleration for pattern search and matching
- Improved BW and Latency

FPGA + Persistent Memory

CPU <-> Coherent / PCIe IF <-> FPGA <-> Persistent Memory

- Unique Competitive **Solution** to address Big, **Fast** and ***Predictable*** Data needs
- Workload offload to FPGA for **Acceleration**
- Memory Semantics
- **Low** Memory **latency access** at **high BW** throughput
- Leverage on different **Memory Hierarchy**
- **Deterministic** high **transaction** per second
- **Coherent** and non-coherent Persistent Memory options (through Coherent Interconnect or PCIe)
- **Cost effective & Flexible** capacity and performance scale-out model

Thank you

sean.in@intel.com

Please attend other Intel sessions
&
Intel Keynote on August 7th @ 11AM