INTEL® OPTANE DC PMM: AN INTRODUCTIONIN

Ramesh Peri, Intel® Corporation

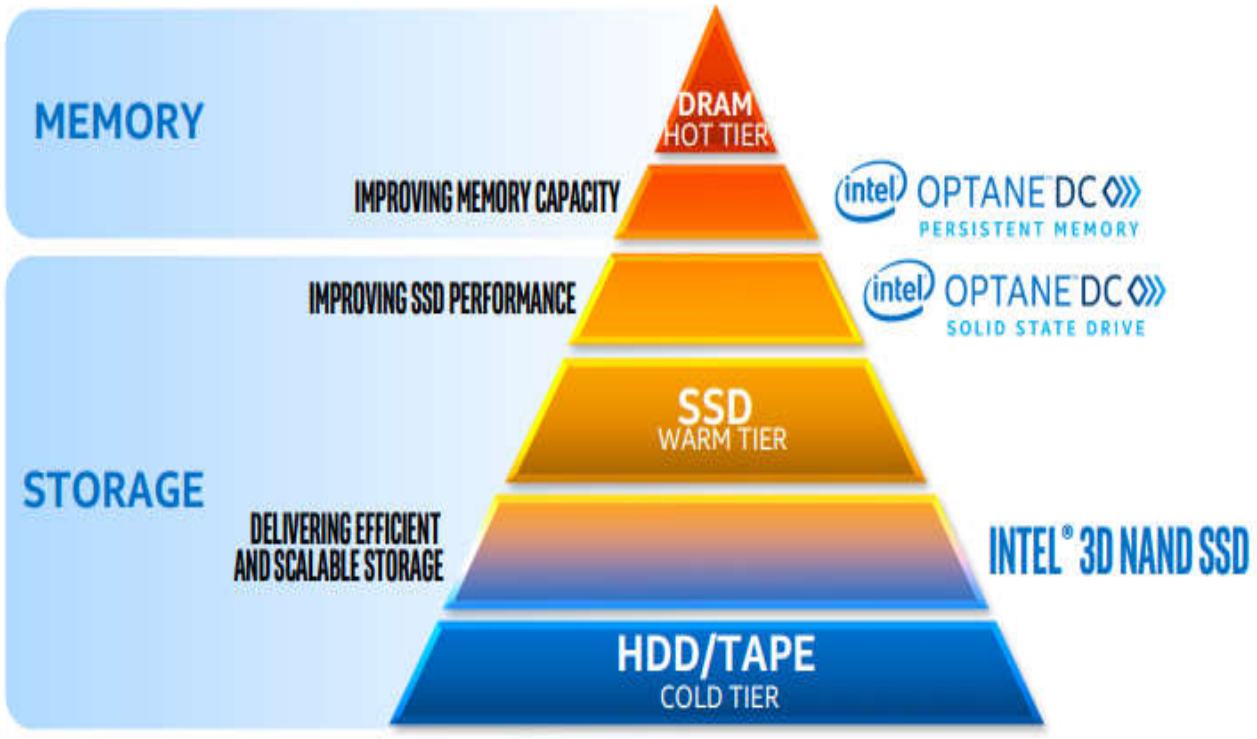
A Question

What if you have a computing platform with infinite compute capacity and infinite amount of memory ?

Agenda

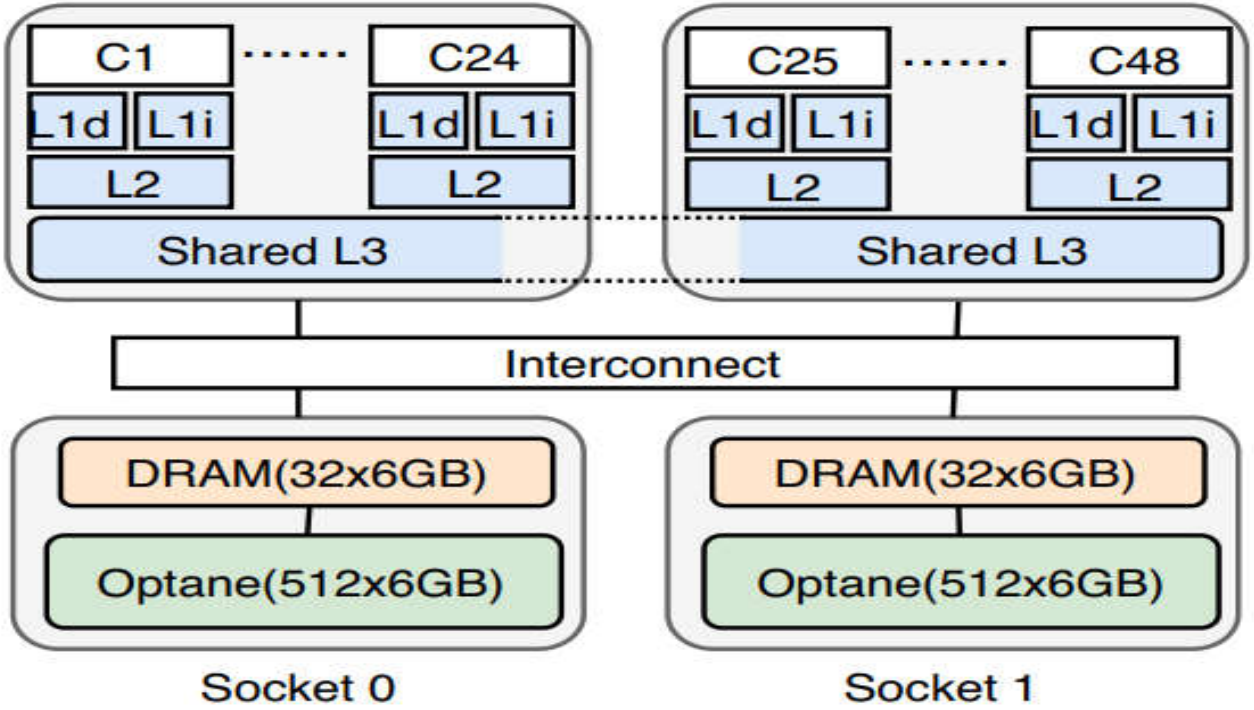
- Introduction to Intel® Optane DC PM
- Introduction to Graph Analytics & Experimental Results
- Tools for Optane DC PMM

New Memory: Intel® Optane DC PM

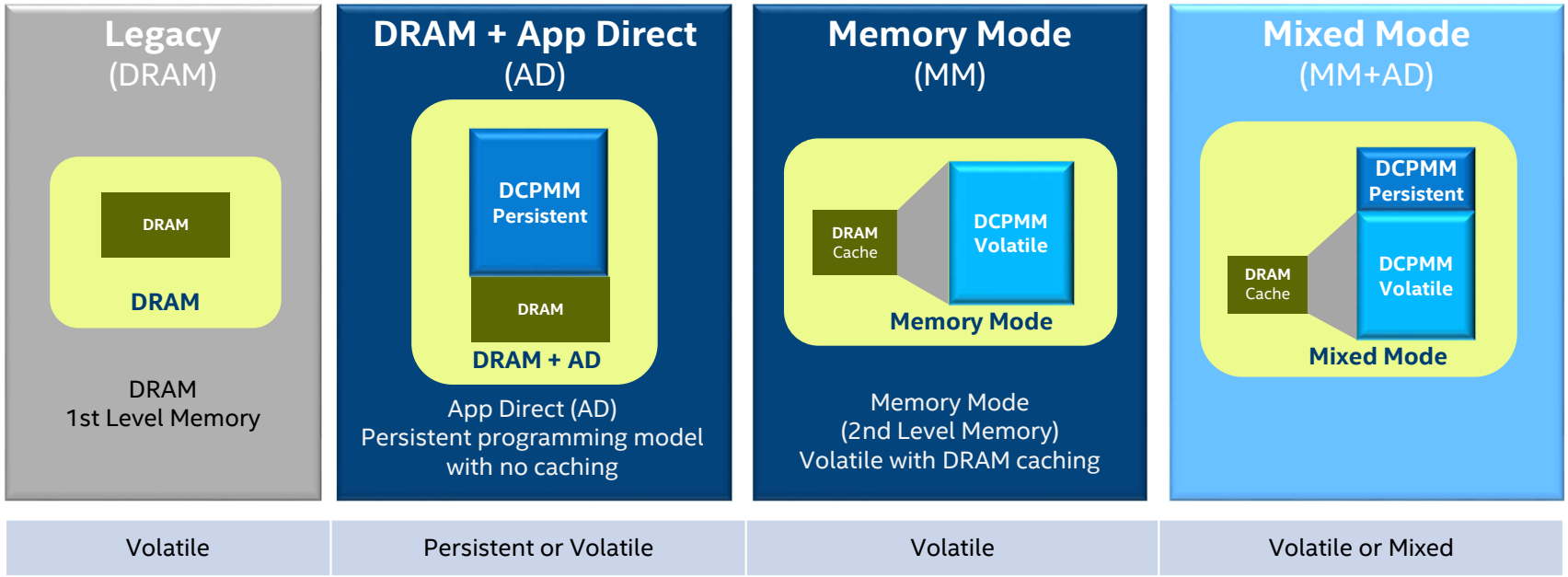


Intel CascadeLake Platform

System Architecture



Programming Models



Choice in configuring the system for best performance based on the application

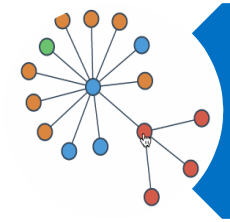
Optane PMM

- Optane is 3x slower for random and 2x slower for sequential read than LDRAM
- Mixing reads/writes hurts Optane performance significantly compared to DRAM
- Bandwidth (GB/s)

	Optane	DRAM
Seq Read	40	105
Random Read	10	70

- https://software.intel.com/en-us/articles/intelr-memory-latency-checker
  - A tool for characterizing memory system behavior
- https://arxiv.org/abs/1903.05714
  - A detailed study of Optane PMM system behavior

GraphAnalytics



Graph Analytics is a good match for Intel Optane DC PMM due to its memory foot print

Can have cost advantage over existing distributed/cluster based solution

Have to pay careful attention to Optimizations in large memory systems to realize full potential

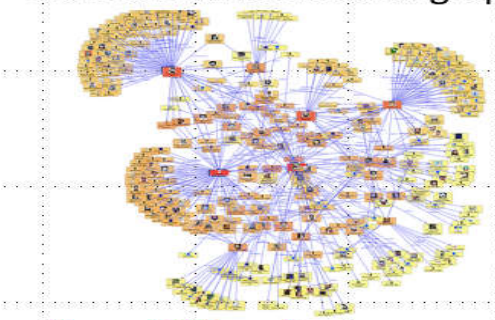
Joint work with UT Austin

Graph Analytics

Applications:
machine learning and network analysis

Credits: Wikipedia, SFL Scientific, MakeUseOf

Datasets: unstructured graphs



Need TBs of memory

Credits: Sentinel Visualizer

BC	• Influential people in social networks, • critical nodes in communication networks
PR	• Importance of webpages
CC	• Finding clusters in social networks, disease propagation
SSSP	• Maps, routing
Kcore	• Hierarchical structure analysis
BFS	• Maps, routing

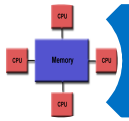
Graph Analytics (cont)

Graphs used in this study

	kron30	clueweb12	uk14	rmat32	wdc12
$\|V\|$	1,073M	978M	788M	4295M	3,563M
$\|E\|$	10,791M	42,574M	47,615M	68,719M	128,736M
$\|E\|/\|V\|$	16	44	60.4	16	36
max D_{out}	3.2M	7,447	16,365	10.4M	55,931
max D_{in}	3.2M	75M	8.6M	10.4M	95M
Approx. diameter	6	498	2498	7	5274
Size on Disk (GB)	136	325	361	544	986

Clueweb12, wdc12, web – webcrawls
uk14– transportation network
Rmat32, kron30 – synthetic graph

Approaches to Graph Analytics



Shared memory systems

Distributed systems

Out of core using secondary storage

Some Graph Systems

Galois, UT Austin

GraphIT, MIT

GraphBlas, Texas A&M

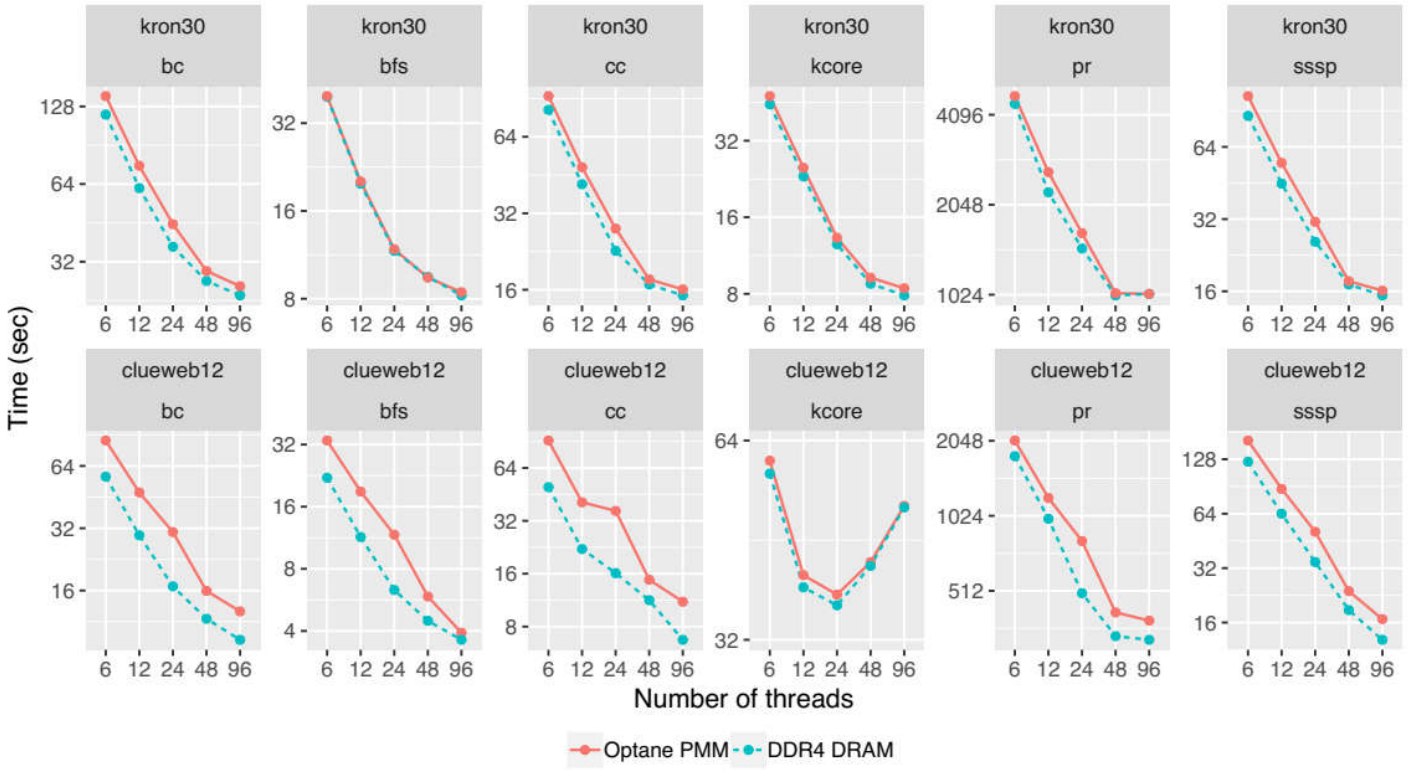
Snap, Stanford

Gunrock, UC Davis

Experimental Setup

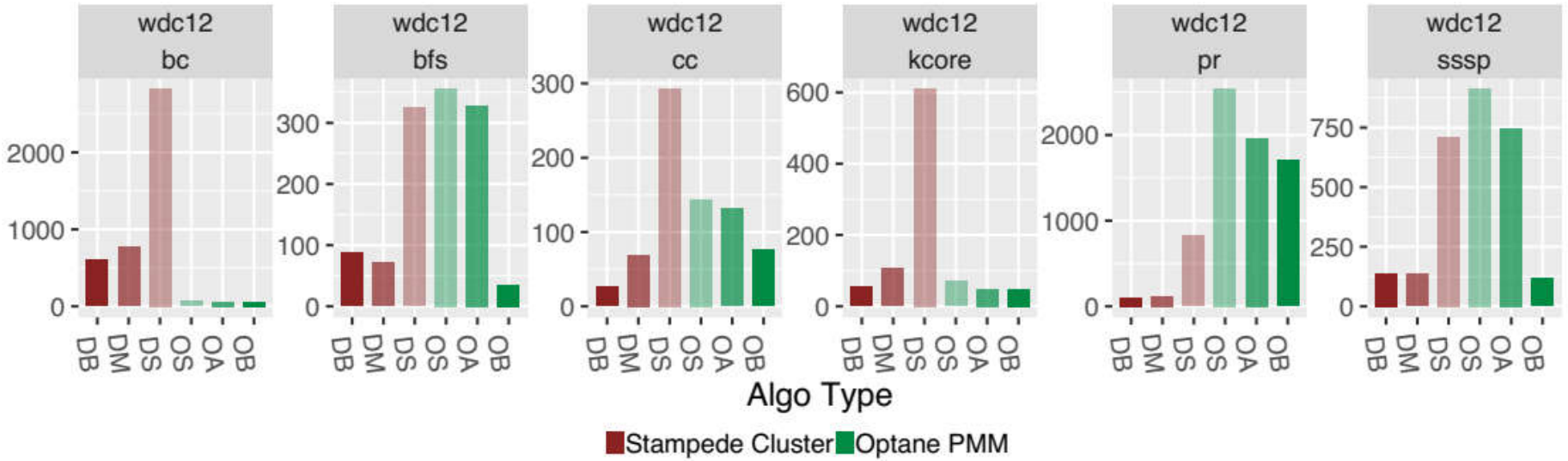
	Optane DC PMM system	DRAM System	Stampede Cluster	DRAM system
Processor	CascadeLake	Cascadelake	Intel Xeon Platinum 8160 ("Skylake") 256 nodes	Intel Xeon Platinum 8180 (Skylake)
Sockets	2	2	2	4
Cores/Socket	24	24	24	28
Threads/Core	2	2	2	2
RAM	384	384	192	1.5TB
Network	NA	NA	100Gb/sec Intel Omni-Path	NA
NVM (Xpoint)	6TB	NA	NA	NA

DRAM vs Optane DC PMM



Optane DC PMM is comparable in performance/scaling to a DRAM only system for datasets that fit in the memory

Distributed Cluster vs Optane DC PMM



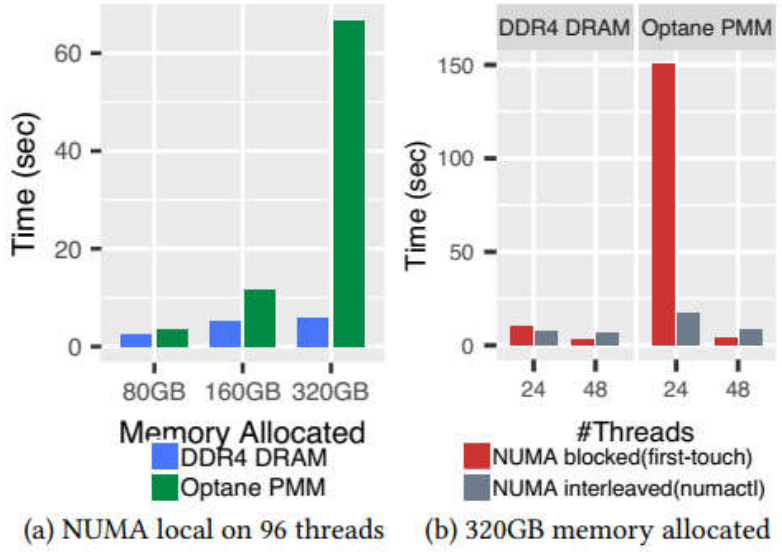
DB – Distributed best (all 256 hosts)
DM – Distributed min (minimum hosts)
DS – Distributed Same (80 threads on min hosts to hold graph)
OS – Optane Same (same algo and threads as DS)
OA – Optane All (same algo as DS, DM, DB)
OB – Optane Best

Single OptaneDC system performance is better than a cluster on 4 out of 6

NUMA Allocation

- Allocation Policies
  - NUMA Local
  - NUMA Interleaved
  - NUMA blocked
- Set by numctl or by the application using anonymous mmap and first touch

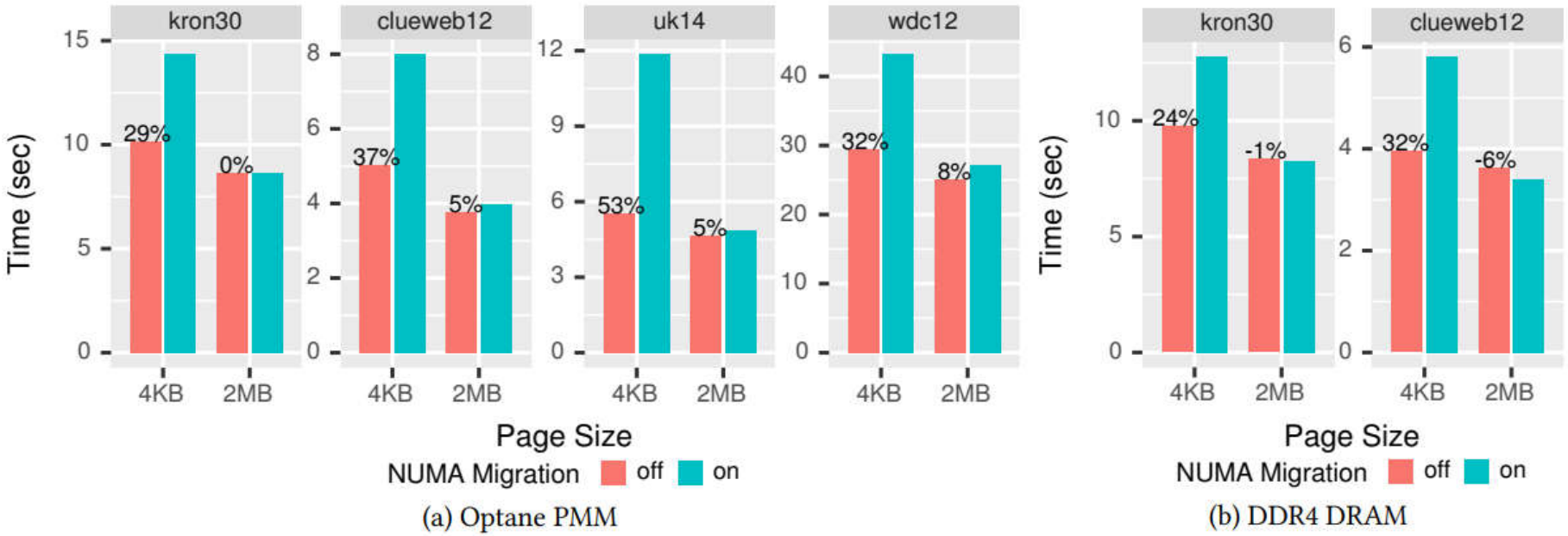
Performance on a micro benchmark



(a) NUMA local on 96 threads (b) 320GB memory allocated

Allocation policy must be carefully selected based on number of threads and working set size

NUMA migration



Cost of page migration on Optane DC is higher compared to DRAM
Large page sizes reduce page migration significantly and hence performance is better

Intel® VTune™ Amplifier - Platform Profiler

Longer runs – system-wide data

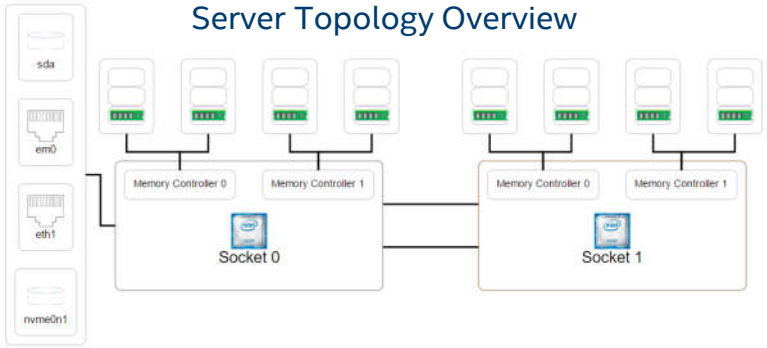
Interactive Topology Diagrams

- System configuration
- Memory channel configurations

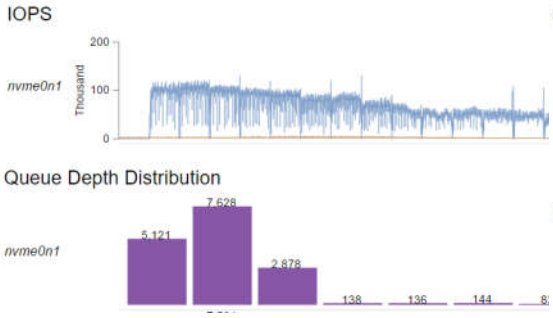
Performance metrics

- Low overhead (targeting < 1%) – coarse grain
- Sampling OS and HW performance counters
- Extended capture (min. to hours)
- Open Data model with RESTful API for easy analysis by scripts

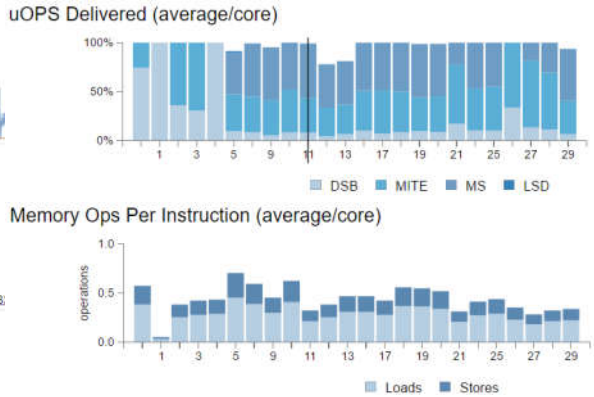
Server Topology Overview



Timelines and Histograms



Core to Core Comparisons

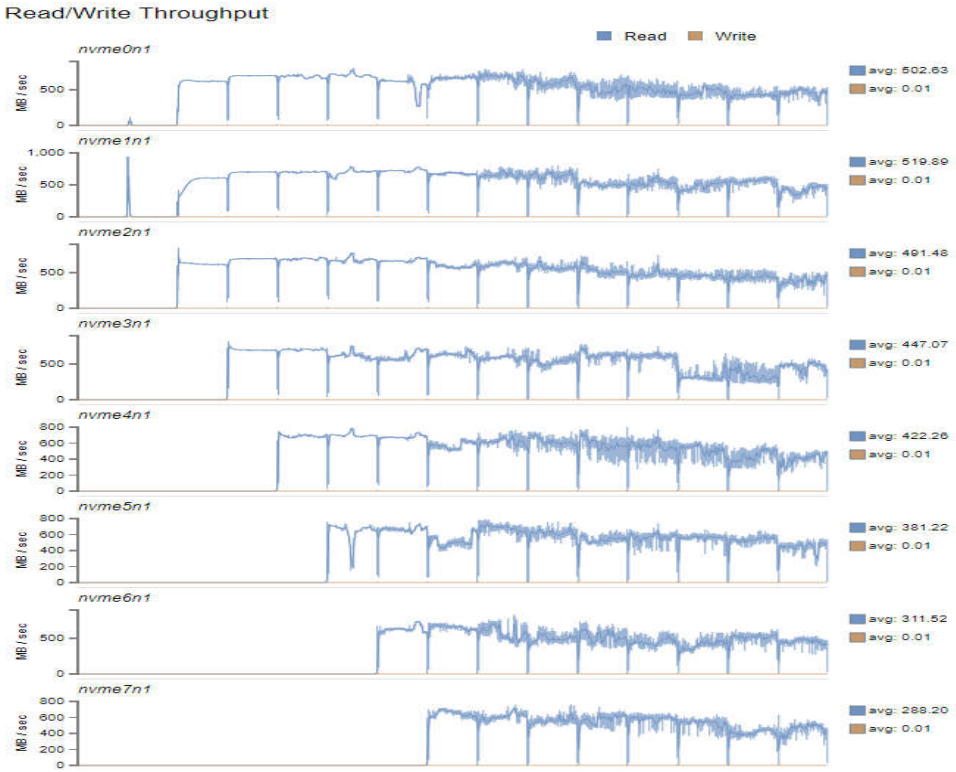


Intel® VTune Amplifier -

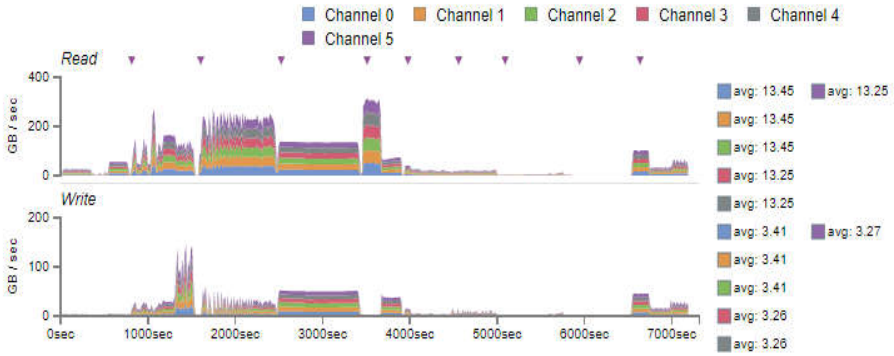
Platform Profiler

Identify utilization imbalances

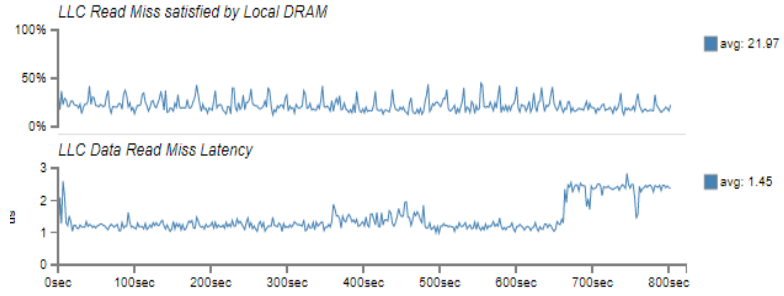
Traffic Patterns



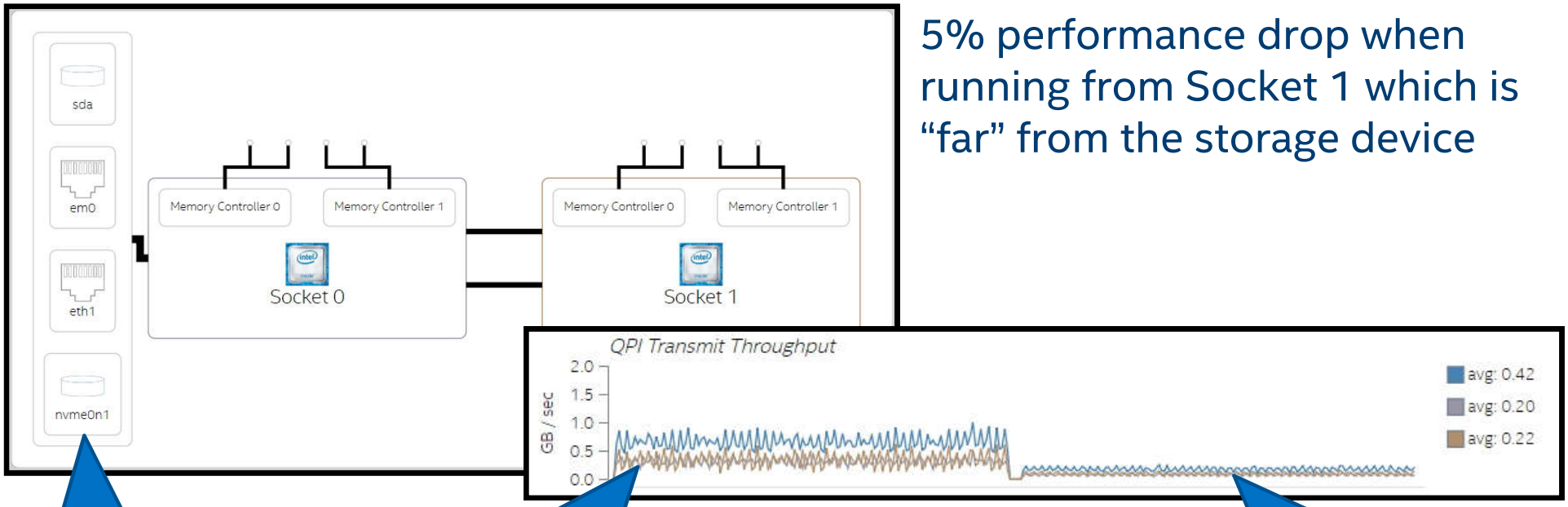
Traffic Patterns



NUMA and Latencies



Configuration Matters



5% performance drop when running from Socket 1 which is "far" from the storage device

Storage device connected to Socket 0

Interconnect traffic from an application running on Socket 1 doing IO

Interconnect traffic from an application running on Socket 0 doing IO

Download the Tools

Software tools for Intel® Optane™ DC Persistent Memory
Free downloads and technical articles

- software.intel.com/persistent-memory/tools

Intel® VTune™ Amplifier – Performance Profiler

- software.intel.com/vtune

Intel® Inspector – Persistence and Thread Debugger

- software.intel.com/inspector

CONCLUSION

- Intel® Optane DC PM is an exciting new memory technology that gives significant boost to memory capacity
- Shared Memory Algorithms can deal with very large datasets and require less hardware and are simpler than Distributed Algorithms
  - Need to be tuned to get good performance
- Applications and OS need to take into the fact that we can now have terbytes of main memory
- Rich set of Performance events available understand and tune the behavior of the processor and Optane DIMMs.
  - Need to provide actionable insight to users through tools

Try Optane DC PM for your large applications
Develop new tools and capabilities to tune applications on this platform

Legal Disclaimer & Optimization Notice

INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS”. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Optimization Notice

Intel’s compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804