Summit at the Oak Ridge Leadership Computing Facility

Judy Hill
Oak Ridge Leadership Computing Facility
Oak Ridge National Laboratory

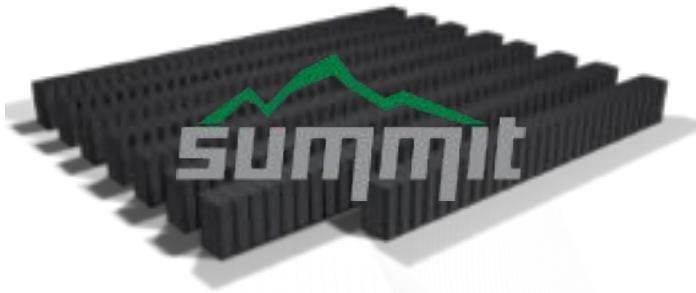
July 30, 2018

Argonne Training Program on Extreme-Scale Computing

ORNL is managed by UT-Battelle
for the US Department of Energy

Outline

- OLCF Roadmap to Exascale
- Summit Architecture Details
- Programming Considerations for Heterogeneous Systems
- Early Summit Application Results



What is the Leadership Computing Facility (LCF)?

- Collaborative DOE Office of Science program at ORNL and ANL
- Mission: Provide the computational and data resources required to solve the most challenging problems.
- 2-centers/2-architectures to address diverse and growing computational needs of the scientific community
- Highly competitive user allocation programs (INCITE, ALCC).
- Projects receive 10x to 100x more resource than at other generally available centers.
- LCF centers partner with users to enable science & engineering breakthroughs (Liaisons, Catalysts).

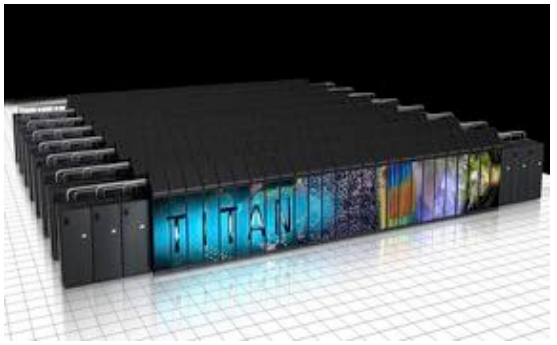
OAK RIDGE National Laboratory
Argonne NATIONAL LABORATORY

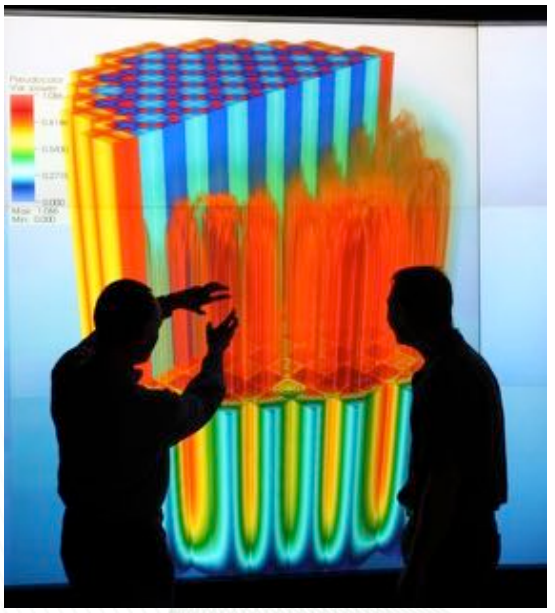
OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Oak Ridge Leadership Computing Facility Mission

The OLCF is a DOE Office of Science National User Facility whose mission is to enable breakthrough science by:

- Fielding the most powerful capability computers for scientific research,
- Building the required infrastructure to facilitate user access to these computers,
- Selecting a few time-sensitive problems of national importance that can take advantage of these systems,
- And partnering with these teams to deliver breakthrough science.



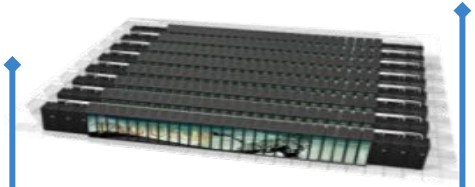


OLCF Path to Exascale

Competitive procurement asking for:

- 50–100× application performance of Titan
- Support for traditional modeling and simulation, high-performance data analysis, and artificial intelligence applications
- Peak performance of at least 1300 PF
- Smooth transition for existing and future applications

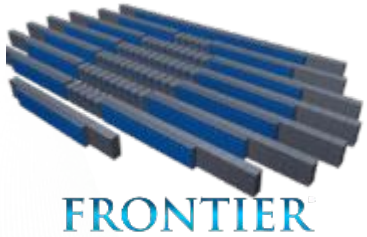
> The Exascale Computing Project has emphasized that Exascale is a measure of application performance, and this RFP reflects that, asking for nominally 50× improvement over Sequoia and Titan.
>
> -- Design Reviewer



2008 — Jaguar: 2.3 PF
World's Fastest

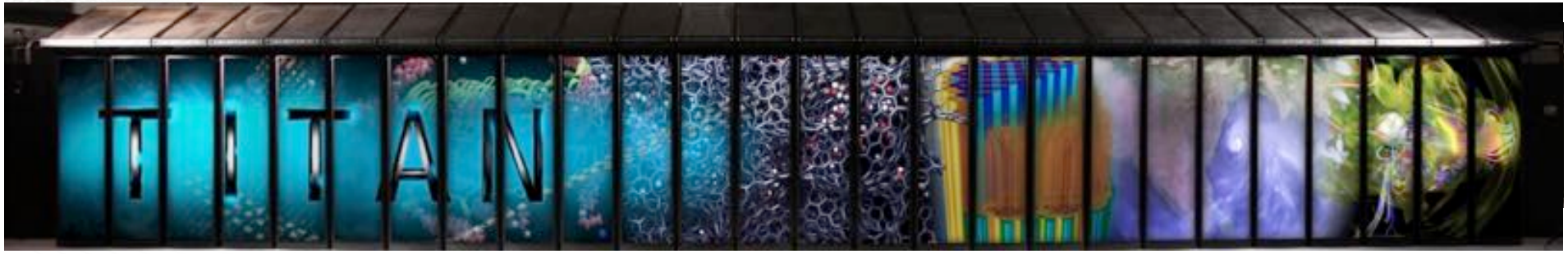
2012 — Titan: 27 PF
Accelerated Computing
World's Fastest

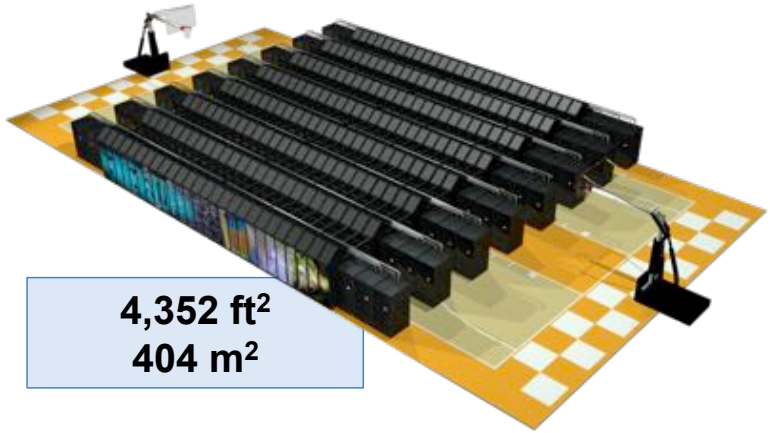
2017 — Summit: 200 PF
Accelerated Computing
5–10× Titan Performance



2021 — Frontier: >1000 PF
Competitive Procurement
5-10× Summit
Performance

ORNL's "Titan" Hybrid System:
Cray XK7 with AMD Opteron and NVIDIA Tesla processors





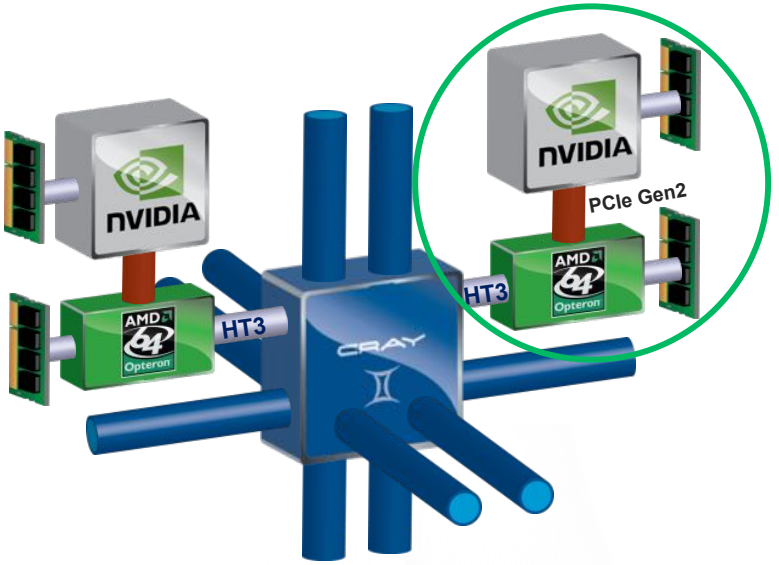
4,352 ft²
404 m²

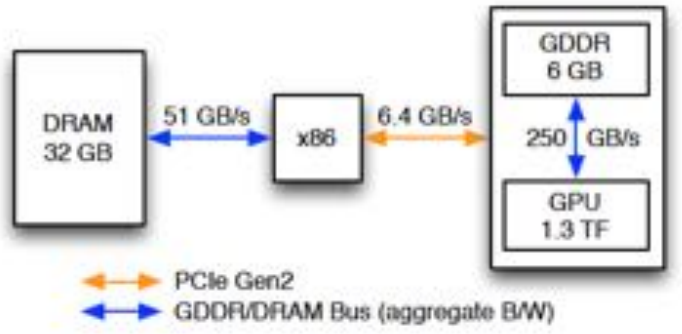
SYSTEM SPECIFICATIONS:

- Peak performance of 27 PF
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - NVIDIA Tesla "K20x" GPU
  - 32 + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 9 MW peak power

Cray XK7 Compute Node

XK7 Compute Node Characteristics
AMD Opteron 6274 “Interlagos” 16 core processor @ 2.2GHz
NVIDIA Tesla K20X “Kepler” 1.31 TF 6GB GDDR5 ECC memory 250 GB/s Memory BW 2688 CUDA cores
Host Memory 32GB 1600 MHz DDR3 ECC memory
Gemini High Speed Interconnect
Four compute nodes per XK6 blade. 24 blades per rack





Coming Soon: Summit is replacing Titan as the OLCF's leadership supercomputer

Summit i al Laboratory's newest s

Coming Soon: Summit is replacing Titan as the OLCF's leadership supercomputer

- Many fewer nodes
- Much more powerful nodes
- Much more memory per node and total system memory
- Faster interconnect
- Much higher bandwidth between CPUs and GPUs
- Much larger and faster file system

Feature	Titan	Summit
Application Performance	Baseline	5-10x Titan
Number of Nodes	18,688	4,608
Node performance	1.4 TF	42 TF
Memory per Node	32 GB DDR3 + 6 GB GDDR5	512 GB DDR4 + 96 GB HBM2
NV memory per Node	0	1600 GB
Total System Memory	710 TB	>10 PB DDR4 + HBM2 + Non-volatile
System Interconnect	Gemini (6.4 GB/s)	Dual Rail EDR-IB (25 GB/s)
Interconnect Topology	3D Torus	Non-blocking Fat Tree
Bi-Section Bandwidth	112 TB/s	115.2 TB/s
Processors	1 AMD Opteron™ 1 NVIDIA Kepler™	2 IBM POWER9™ 6 NVIDIA Volta™
File System	32 PB, 1 TB/s, Lustre®	250 PB, 2.5 TB/s, GPFS™
Power Consumption	9 MW	13 MW

Summit Overview

OpenPOWER™

Compute Node

2 x POWER9
6 x NVIDIA GV100
NVMe-compatible PCIe 1600 GB SSD

25 GB/s EDR IB- (2 ports)
512 GB DRAM- (DDR4)
96 GB HBM- (3D Stacked)
Coherent Shared Memory

Compute Rack

18 Compute Servers
Warm water (70°F direct-cooled components)
RDHX for air-cooled components

39.7 TB Memory/rack
55 KW max power/rack

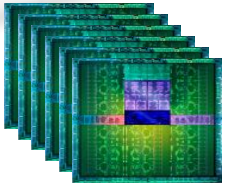
Compute System

10.2 PB Total Memory
256 compute racks
4,608 compute nodes
Mellanox EDR IB fabric
200 PFLOPS
~13 MW

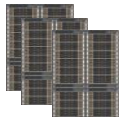
Components

IBM POWER9
- 22 Cores
- 4 Threads/core
- NVLink

NVIDIA GV100
- 7 TF
- 16 GB @ 0.9 TB/s
- NVLink



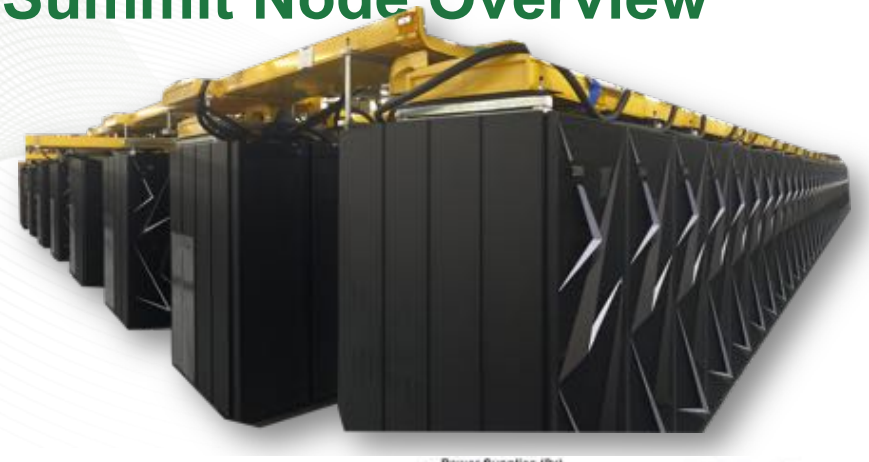
GPFS File System

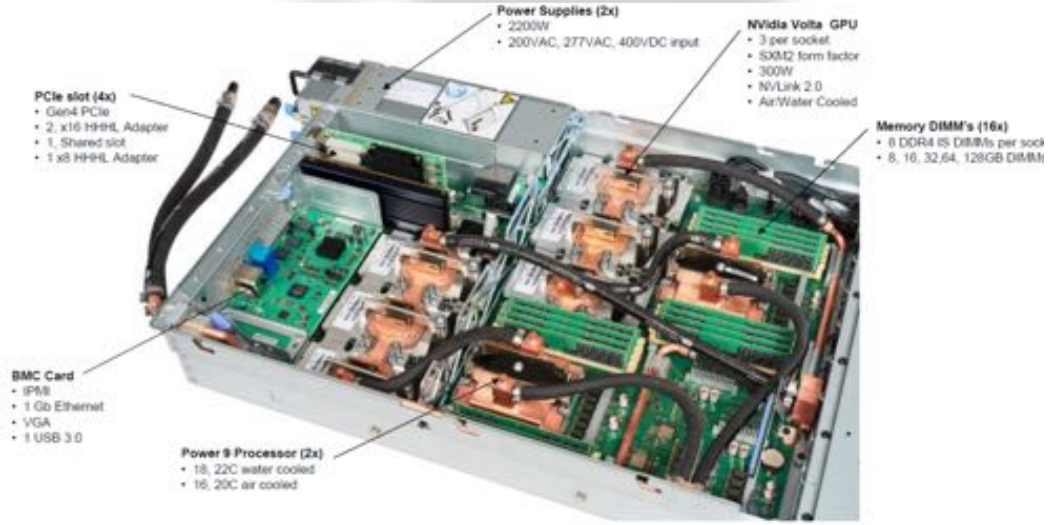


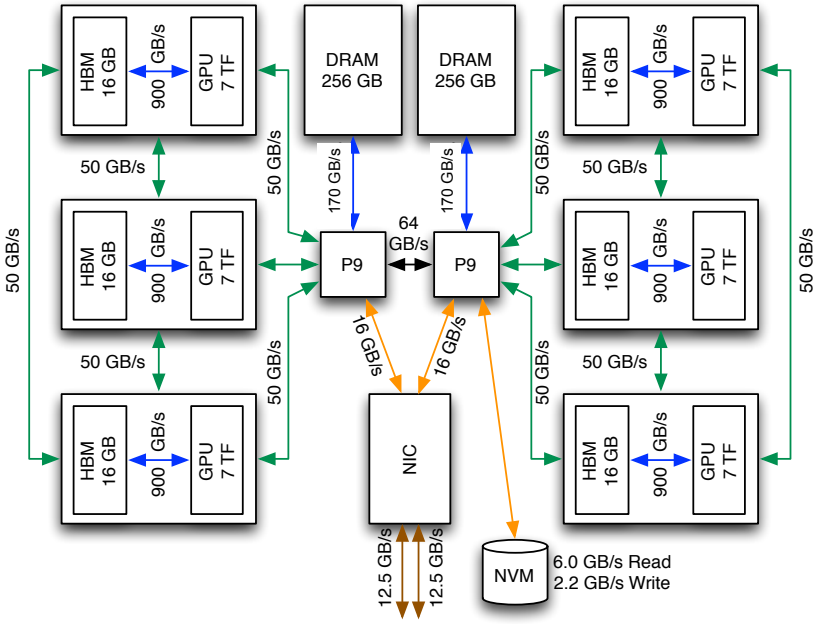
250 PB storage
2.5 TB/s read, 2.5 TB/s write
(**2.5 TB/s sequential and 2.2 TB/s random I/O)

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Summit Node Overview







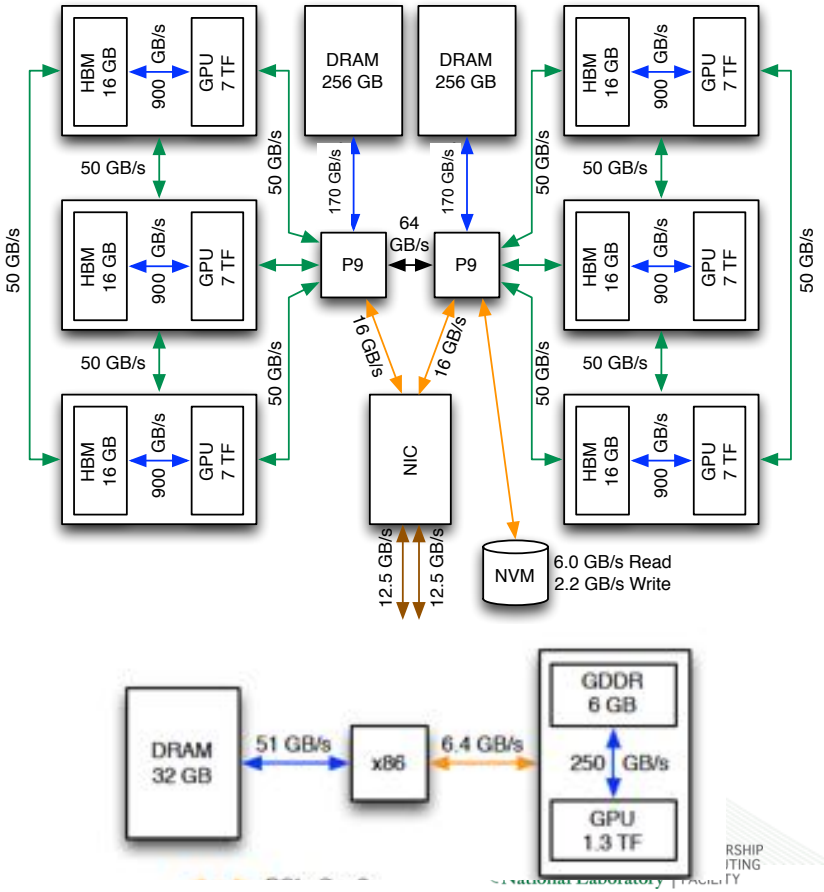
TF	42 TF (6x7 TF)
HBM	96 GB (6x16 GB)
DRAM	512 GB (2x16x16 GB)
NET	25 GB/s (2x12.5 GB/s)
MMsg/s	83

HBM/DRAM Bus (aggregate B/W)
NVLINK
X-Bus (SMP)
PCIe Gen4
EDR IB

HBM & DRAM speeds are aggregate (Read+Write).
All other speeds (X-Bus, NVLink, PCIe, IB) are bi-directional.

Summit Node Overview: System Balance Ratios

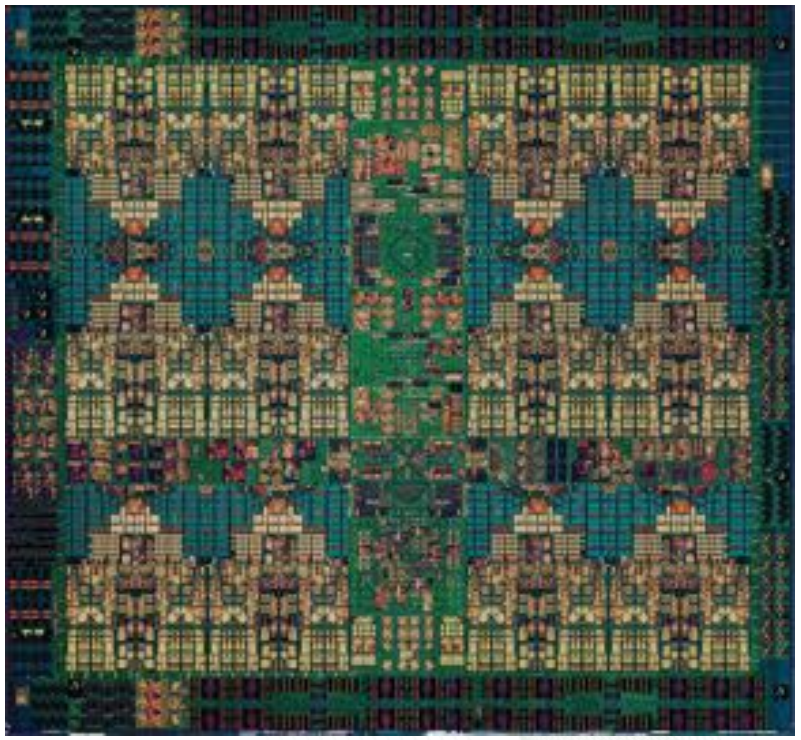
	Summit	Titan
Memory subsystem to Intra-node connectivity ratios		
HBM BW : DDR BW	15.8	4.9
HBM BW : CPU-GPU BW	18	39
Per HBM BW : GPU-GPU BW	18	--
DDR BW : CPU-GPU BW	1.13	8
HBM capacity : GPU-GPU BW	0.32	--
Memory subsystem to FLOPS ratios		
Memory capacity : GFLOPS	0.01	0.03
Interconnect subsystem to FLOPS ratios		
Injection BW : GFLOPS	0.0006	0.004
Other ratios		
Filesystem : Memory capacity	89	42
FLOPS : Power (MW)	15.4	3



Reference: Vazhkudai, et. al. The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems. SC18 Proceedings. To appear.

IBM Power9 Processor

- Up to 24 cores
  - CORAL has 22 cores for yield optimization on first processors
- PCI-Express 4.0
  - Twice as fast as PCIe 3.0
- NVLink 2.0
  - Coherent, high-bandwidth links to GPUs
- 14nm FinFET SOI technology
  - 8 billion transistors
- Cache
  - L1I: 32 KiB per core, 8-way set associative
  - L1D: 32KiB per core, 8-way
  - L2: 258 KiB per core
  - L3: 120 MiB eDRAM, 20-way



Stream benchmark: Summit vs Titan

- A simple synthetic benchmark program that measures achievable memory bandwidth (in GB/s) under OpenMP threading.

System Cores	Peak (Summit) 44	Titan 16
Copy	274.6	34.9
Scale	271.4	35.3
Add	270.6	33.6
Triad	275.3	33.7
Peak (theoretical)	340	51.2
Fraction of Peak	82%	67%

DRAM Bandwidth

System	Peak (Summit)	Titan
Copy	789	181
Scale	788	181
Add	831	180
Triad	831	180
Peak (theoretical)	900	250
Fraction of Peak	92%	72%

GDDR Bandwidth

For Peak (Summit):
- GCC compiler
- Best result in 1000 tests
- Runtime variability up to 9%

Slide courtesy of Wayne Joubert, ORNL

NVIDIA Volta Details

	Tesla V100 for NVLink	Tesla V100 for PCIe
PERFORMANCE with NVIDIA GPU Boost™	DOUBLE-PRECISION 7.8 TeraFLOPS	DOUBLE-PRECISION 7 TeraFLOPS
	SINGLE-PRECISION 15.7 TeraFLOPS	SINGLE-PRECISION 14 TeraFLOPS
	DEEP LEARNING 125 TeraFLOPS	DEEP LEARNING 112 TeraFLOPS
INTERCONNECT BANDWIDTH Bi-Directional	NVLINK 300 GB/s	PCIE 32 GB/s
MEMORY CoWoS Stacked HBM2	CAPACITY 16 GB HBM2	
	BANDWIDTH 900 GB/s	

TensorCores™
Mixed Precision
(16b Matrix-Multiply-Add and 32b Accumulate)

Note: The performance numbers are peak and not representative of Summit's Volta

NVLink Bandwidth

- Measured from core 0 the achieved CPU-GPU NVLink rates with a modified bandwidthTest from NVIDIA CUDA Samples

GPU	0	1	2	3	4	5	peak
Host to Device	45.93	45.92	45.92	40.63	40.59	40.64	50
Device to Host	45.95	45.95	45.95	36.60	36.52	35.00	50
Bi-Directional	86.27	85.83	77.36	66.14	65.84	64.76	100

Single Node Single GPU NVLink Rates (GB/s)

- Not necessarily a use case that most applications will employ

Slide courtesy of Wayne Joubert, ORNL

NVLink Bandwidth

- Measured the achieved CPU-GPU NVLink rates with a modified bandwidthTest from NVIDIA CUDA Samples using multiple MPI process evenly spread between the sockets.

MPI Process Count	1	2	3	4	5	6	Peak (6)
Host to Device	45.93	91.85	137.69	183.54	229.18	274.82	300
Device to Host	45.95	91.90	137.85	183.80	225.64	268.05	300
Bi-Directional	85.60	172.59	223.54	276.34	277.39	278.07	600

NVLink Rates with MPI Processes (GB/s)

- Ultimately limited by the CPU memory bandwith
- 6 ranks driving 6 GPUs is an expected use case for many applications

Slide courtesy of Wayne Joubert, ORNL

NVLink Bandwidth

- Measured the achived NVLink transfer rates between GPUs, both within a socket and across them, using p2pBandwidthLatencyTest from NVIDIA CUDA Samples. (Peer-to-Peer communication turned on).

Socket	0	1	Cross	Peak
Uni-Directional	46.33	46.55	25.89	50
Bi-Directional	93.02	93.11	21.63	100

NVLink Rates for GPU-GPU Transfers (GB/s)

- Cross-socket bandwidth is much lower than that between GPUs attached to the same CPU socket

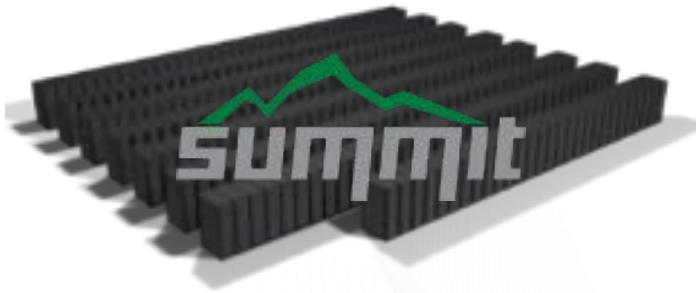
Slide courtesy of Wayne Joubert, ORNL

Summit is still in pre-production

- We expect to accept the machine in Fall of 2018, allow early users on this year, and allocate our first users through the INCITE program in January 2019.

Outline

- OLCF Roadmap to Exascale
- Summit Architecture Details
- Programming Considerations for Heterogeneous Systems
- Early Summit Application Results



Summit Programming Environment

- Compilers supporting OpenMP and OpenACC
  - IBM XL, PGI, LLVM, GNU, NVIDIA
- Libraries
  - IBM Engineering and Scientific Subroutine Library (ESSL)
  - FFTW, ScaLAPACK, PETSc, Trilinos, BLAS-1,-2,-3, NVBLAS
  - cuFFT, cuSPARSE, cuRAND, NPP, Thrust
- Debugging
  - Allinea DDT, IBM Parallel Environment Runtime Edition (pdb)
  - Cuda-gdb, Cuda-memcheck, valgrind, memcheck, helgrind, stacktrace
- Profiling
  - IBM Parallel Environment Developer Edition (HPC Toolkit)
  - VAMPIR, Tau, Open|Speedshop, nvprof, gprof, Rice HPCToolkit

Summit vs Titan PE comparison

Compiler	Titan	Summit
PGI	Yes	Yes
GCC	Yes	Yes
XL	No	Yes
LLVM	No	Yes
Cray	Yes	No
Intel	Yes	No

Debugger	Titan	Summit
DDT	Yes	Yes
cuda-gdb, -memcheck	Yes	Yes
Valgrind, memcheck, helgrind	Yes	Yes
Stack trace analysis tool	Yes	Yes
pdb	No	Yes

Performance Tools	Titan	Summit
Open|SpeedShop	Yes	Yes
TAU	Yes	Yes
CrayPAT	Yes	No
Reveal	Yes	No
HPCToolkit (IBM)	No	Yes
HPCToolkit (Rice)	Yes	Yes
VAMPIR	Yes	Yes
nvprof	Yes	Yes
gprof	Yes	Yes

The majority of tools available on Titan are also available on Summit. A few transitions may be necessary.

More Information on Compilers

On Titan

- Unaccelerated code
  - PGI, GCC, Cray, Intel
- OpenMP code
  - PGI, GCC, Cray, Intel
- OpenACC code
  - PGI, Cray, GCC (under development)
- CUDA code
  - PGI, GCC, Intel
- CUDA Fortran code
  - PGI

On Summit

- Unaccelerated code
  - PGI, GCC, XL, LLVM
- OpenMP code
  - PGI, GCC, XL, LLVM
- OpenACC code
  - PGI, GCC (under development)
- CUDA code
  - PGI, GCC
- CUDA Fortran code
  - PGI, XL

Users of compilers on Titan that are not available on Summit will require transitioning to one of the supported compilers.

Summit vs Titan Math Library Comparison

Library	OSS or Proprietary	CPU Node	CPU Parallel	GPU
IBM ESSL	Proprietary	✓		✓
FFTW	OSS	✓		✓
ScaLAPACK	OSS		✓	
PETSc	OSS		✓	✓*
Trilinos	OSS		✓	✓*
BLAS-1, -2, -3	Proprietary (thru ESSL)			✓
NVBLAS	Proprietary			✓
cuFFT	Proprietary			✓
cuSPARSE	Proprietary			✓
cuRAND	Proprietary			✓
NPP	Proprietary			✓
Thrust	Proprietary			✓

* Available GPU capabilities in the OSS library will be optimized.

Supporting Other Programming Models

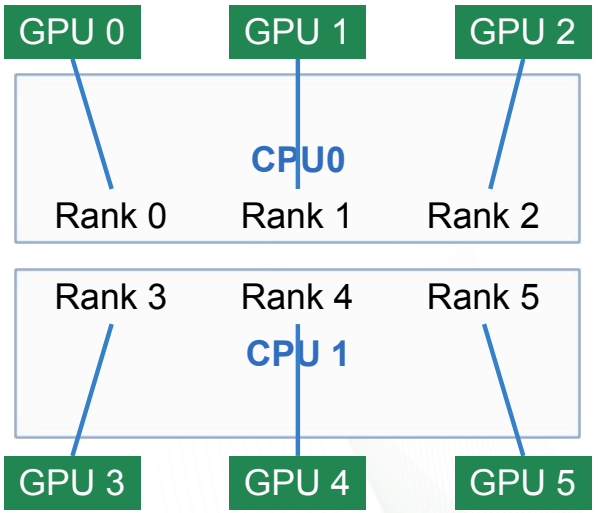
- IBM PAMI (low-level communications interface) will support a variety of established non-MPI programming models
  - Global Arrays (GA) and Aggregate Remote Memory Copy Interface (ARMCI)
  - Charm++
  - GASNet
  - OpenSHMEM
- These programming models are similar to research directions for exascale
- Will support accelerators and node-local NVRAM as communication endpoints

Programming Multiple GPUs

- Multiple paths, with different levels of flexibility and sophistication
  - Simple model looks like Titan
  - Additional models expose the node-level parallelism mode directly
  - Low-level approaches are available, but not what we would recommend to users unless there is a particular reason
- Exposing more (node-level) parallelism is key to scalable applications from petascale up

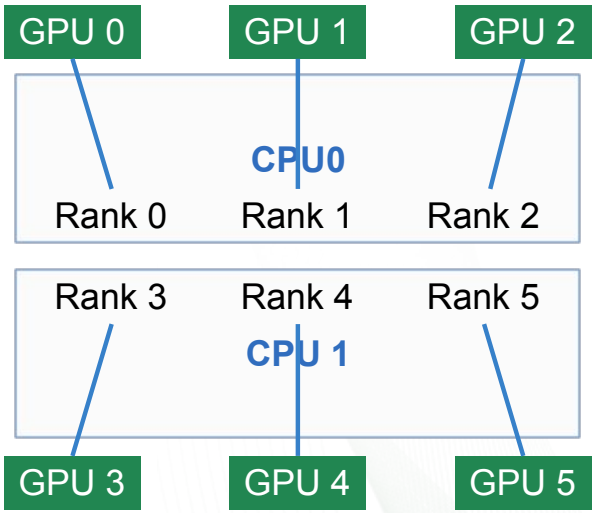
One GPU Per MPI Rank

- Deploy one MPI rank per GPU (6 per node)
  - Bind each rank to a specific GPU
- This model looks like Titan
- MPI ranks can use OpenMP (or pthreads) to utilize more of the CPU cores
  - CPU is only a small percentage of the total FLOPS



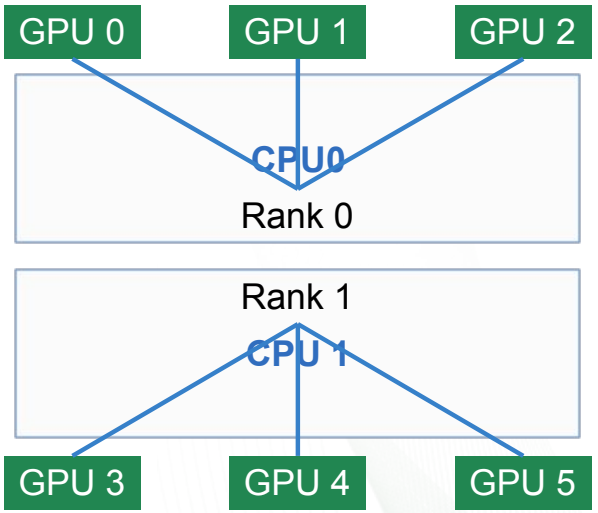
One GPU Per MPI Rank

- Expect this to be the most commonly used approach.
- Pros:
  - ✓ Straightforward extension for those already using Titan
- Cons:
  - Assumes similar amount of work to be done by all ranks
  - Potentially leaves a core on the Power9 unoccupied (or available to do something else)



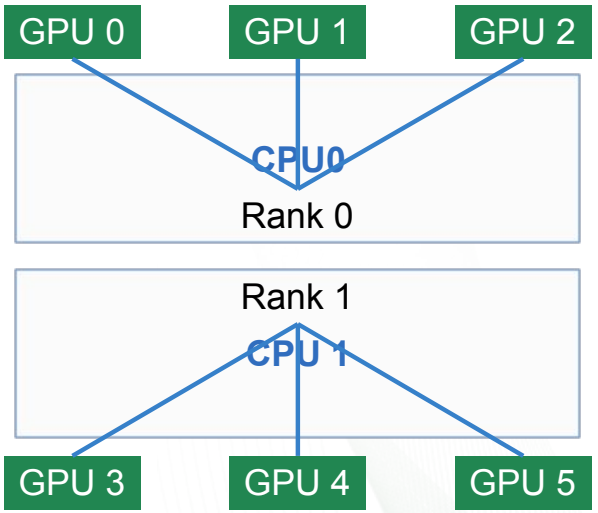
Multiple GPUs Per MPI Rank

- Deploy one MPI rank per 2-6 GPUs
  - Likely configurations:
    - 3 ranks/node (1:2)
    - 2 ranks/node (1:3)
    - 1 rank/node (1:6)
- Use threads and/or language constructs to offload to specific devices
- Multiple approaches possible, depending on language



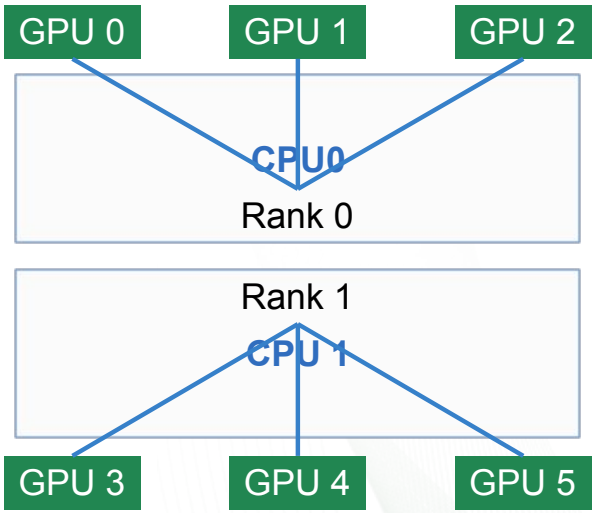
Multiple GPUs Per MPI Rank, Explicit Control

- OpenMP+OpenACC
  - Launch one OpenMP thread per GPU
  - Within each thread make OpenACC calls using `acc_set_device_num()`
- OpenMP 4 (accelerator target)
  - `device_num()` clause
- OpenACC
  - `acc_set_device_num()`
  - (Need to add similar clause for directives)
  - Eventually: compiler+runtime could break up large offload tasks across multiple GPUs automatically
- CUDA
  - `cudaSetDevice()` method



Multiple GPUs Per MPI Rank, Implicit Control

- OpenMP and OpenACC
  - Eventually: compiler+runtime could break up large offload tasks across multiple GPUs automatically
- Task-based execution models are available for CUDA, OpenMP and under development for OpenACC
  - Provide more flexibility to distribute work to multiple GPUs
- Multi-GPU aware libraries
  - CUBLAS
  - CUFFT

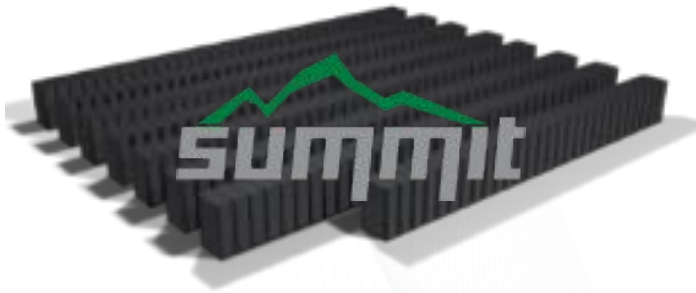


Application Portability

- Portability has been a concern since the beginning of the Titan project
  - Accelerators vs multi-core node architecture
- OLCF has been advocating directive-based programming approaches, and helping to develop them
  - Base code written for multi-core architecture
  - Directives to selectively offload computation to accelerator
  - OpenACC is the primary embodiment of this approach
  - Conceptual spin-off of OpenMP and planned to merge back into OpenMP in time
- CORAL/NERSC-8 systems don't fundamentally change this
  - More GPU accelerators or more cores per node
  - But flattening of memory access times will greatly increase programmability of both architectures

Outline

- OLCF Roadmap to Exascale
- Summit Architecture Details
- Programming Considerations for Heterogeneous Systems
- Early Summit Application Results

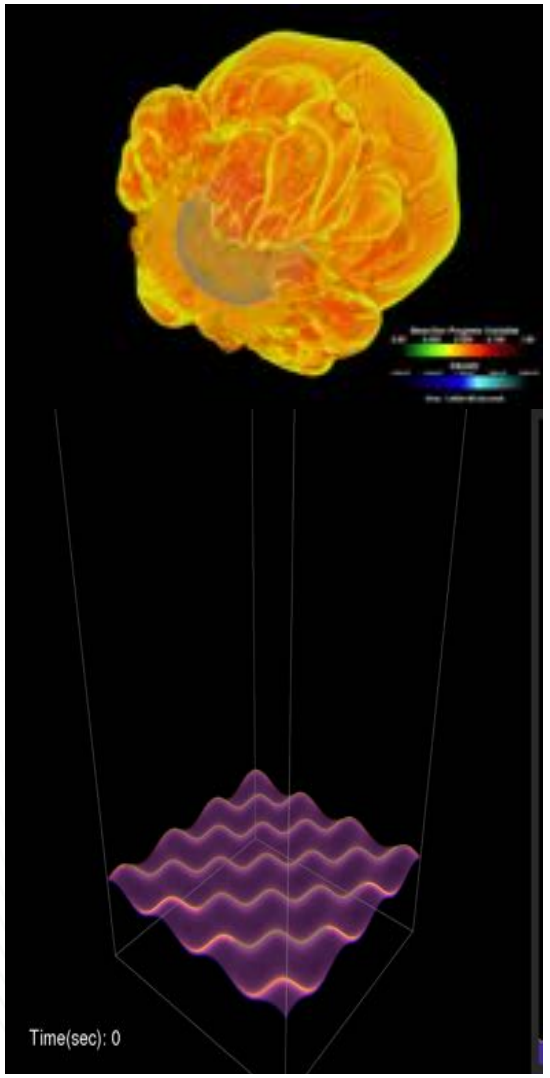


FLASH

- FLASH is a publicly available, component-based, massively parallel, adaptive mesh refinement (AMR) code that has been used on a variety of parallel platforms.
- The code has been used to simulate a variety of phenomenon, including thermonuclear and core-collapse supernovae, galaxy cluster formation, classical novae, the formation of proto-planetary disks, and high-energy-density physics. FLASH's multi-physics and AMR capabilities make it an ideal numerical laboratory for investigations of nucleosynthesis in supernovae.

Targeted for CAAR

1. **Nuclear kinetics (burn unit) threading and vectorization, including Jacobian formation and solution using GPU-enabled libraries**
2. Equation of State (EOS) threading and vectorization
3. Hydrodynamics module performance



Slide courtesy of Bronson Messer, ORNL

FLASH Early Summit Results

- FLASH: Component-based, massively parallel, adaptive-mesh refinement code
  - Widely used in astrophysics community (>1100 publications from >600 scientists)
- CAAR work primarily concerned with increasing physical fidelity by accelerating the nuclear burning module and associated load balancing.
- Summit GPU performance **fundamentally changes the potential science impact** by enabling large-network (i.e. 160 or more nuclear species) simulations.
  - Heaviest elements in the Universe are made in neutron-rich environments – small networks are incapable of tracking these neutron-rich nuclei
  - Opens up the possibility of producing precision nucleosynthesis predictions to compare to observations
  - Provides detailed information regarding most astrophysically important nuclear reactions to be measured at FRIB

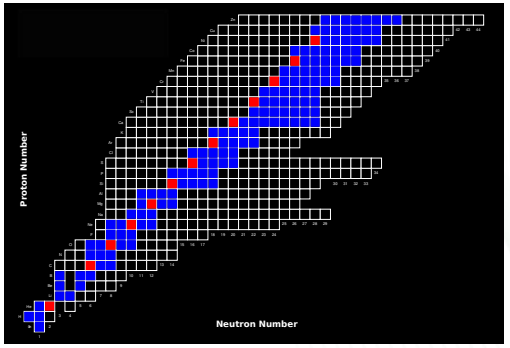
Preliminary results on Summit

GPU+CPU vs. CPU-only performance on Summit for 288-species network : **2.9x**

P9: 24.65 seconds/step
P9 + Volta: 8.5 seconds/step

288-species impossible to run on Titan

NASA, ESA, J. Hester and A. Loll (Arizona St. Univ.)



*Time for 160-species (blue) run on Summit **roughly equal** to 13-species "alpha" (red) network run on Titan*

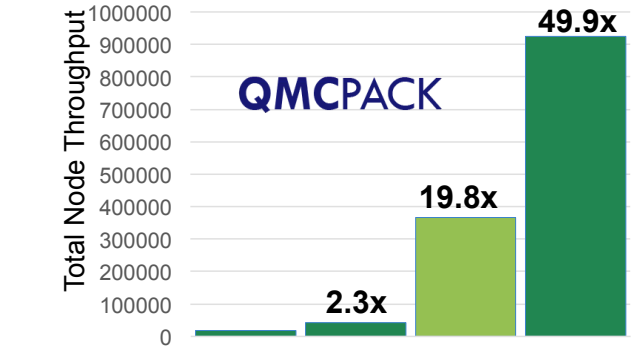
>100x the computation for identical cost

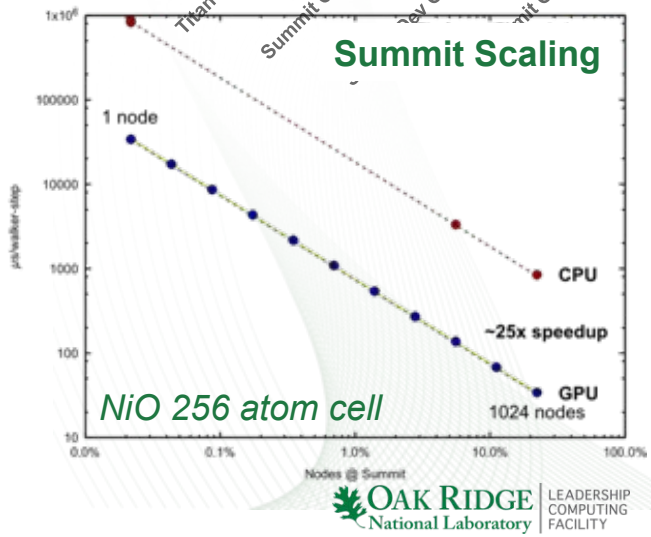
Slide courtesy of Bronson Messer, ORNL

QMCPACK on Summit

- QMCPACK: Accurate quantum mechanics based simulation of materials, including high temperature superconductors.
- QMCPACK runs correctly and with good initial performance on up to 1024 nodes (>20% Summit)
- A Summit node is 50-times faster than a Titan node for this problem, indicating a ~3.7x increase in the complexity of materials (electron count) computable in the same walltime as Titan.
- Summit exceeds performance gains expected based on peak flops by a factor of 1.57x
- New developments for even better Summit utilization:
  - Delayed updates increase compute intensity on GPUs (Blas-2 → Blas-3)
  - Spline buffer (95% of static memory) splitting over multiple GPUs can be used to increase compute density (more walkers) or to enable currently impossibly large systems (6 x 16GB = 96 GB of effective GPU memory)

QMCPACK v3.4.0 NiO 128 atom cell. Power CPU reference uses 2 MPI tasks, 42 OpenMP threads each and optimized "SoA" version.



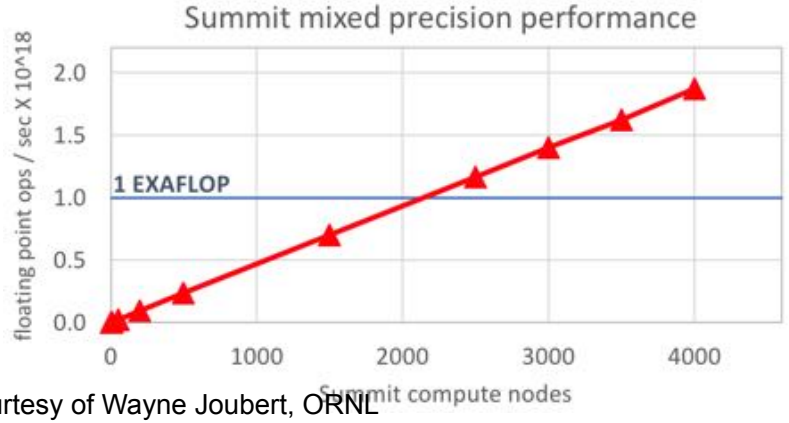


Slide courtesy of Andreas Tillack, ORNL

CoMet: ExaOp Comparative Genomics on Summit

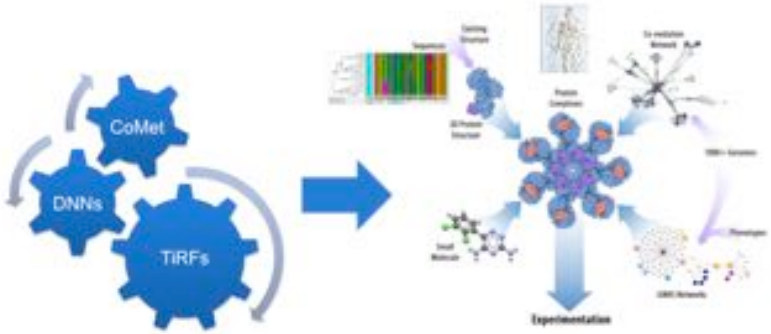
Dan Jacobson, Wayne Joubert (ORNL)

- Modified 2-way CCC algorithm uses NVIDIA Volta Tensor Cores and cuBLAS library to compute counts of bit values
- Near-ideal weak scaling to 4000 nodes (87% of Summit) – **1.8 EF** mixed precision performance reached; 234 quadrillion element comparisons / sec attained
- **4.5X faster** than previous optimized bitwise CCC/sp code on Summit
- **80 TF** mixed precision achieved per GPU for full algorithm – cuBLAS performance per GPU nearly **100 TF**
- Expect **2+ EF mixed precision achievable** on full Summit system



Slide courtesy of Wayne Joubert, ORNL

- **Summit allows us to:**
  - Discover co-evolutionary relationships across a population of genomes at an unprecedented scale
  - Discover epistatic interactions for Opioid Addiction
    - Gordon Bell Prize submission



W. Joubert, J. Nance, S. Climer, D. Weighill, D. Jacobson, "Parallel Accelerated Custom Correlation Coefficient Calculations for Genomics Applications," arxiv 1705.08213 [cs], *Parallel Computing,* accepted.

Acknowledgements

- Entire OLCF Team, specifically OLCF Scientific Computing Group
  - Wayne Joubert, Bronson Messer, Andreas Tillack, Dmitry Liakh, Mark Berrill, Reuben Budiardjia, Matt Norman, and many others

- Additional Reference:
  - Vazhkudai, *et. al.* The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems. SC18 Proceedings. To appear.
- For more Summit info, OLCF will be hosting Summit training, tentatively scheduled for December and February.

https://www.olcf.ornl.gov/for-users/training/training-calendar/

This work was performed under the auspices of the U.S. DOE by Oak Ridge Leadership Computing Facility at ORNL under contracts DEAC05-00OR22725


Questions?