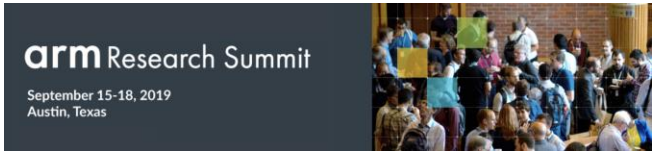
ARM RESEARCH SUMMIT 2019



Argonne NATIONAL LABORATORY

EARLY EXPERIENCE OF THE ARM-BASED HIGH-PERFORMANCE COMPUTING ECO-SYSTEM

VITALI MOROZOV, JAEHYUK KWACK, THOMAS APPLENCOURT, COLLEEN BERTONI, YASAMAN GHADAR, HUIHUO ZHENG, CHRISTOPHER KNIGHT, AND SCOTT PARKER

Argonne Leadership Computing Facility
Argonne National Laboratory

September 16, 2019, Austin, TX

BENCHMARKS & MICRO-ARCHITECTURE CHARACTERISTICS

ASTRA - SANDIA'S NNSA/ASC ADVANCED ARM PROTOTYPE SUPERCOMPUTER

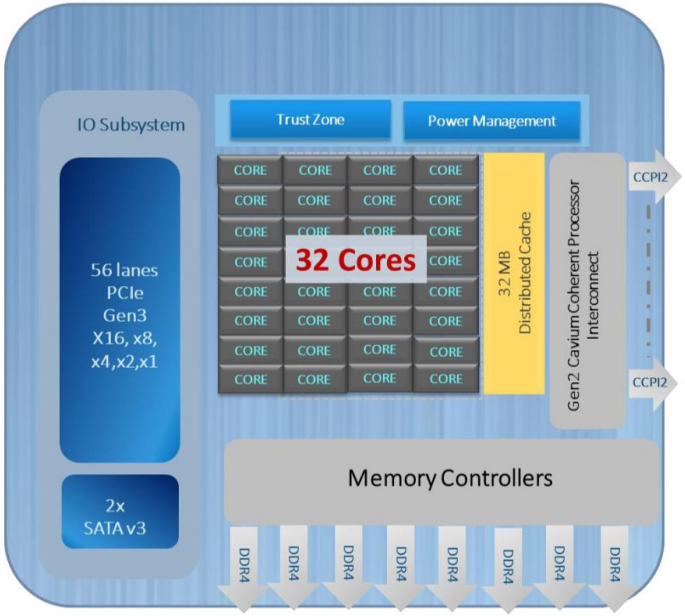
- Goals
  - Evaluate and mature Arm-processor technology for future US capability and capacity systems
  - Expand US supercomputing ecosystem to include HPE and Arm/Cavium/Marvell
  - Develop a full-featured, production-ready open software stack for Arm HPC
- Architecture (2.3 PFLOPs system peak)
  - 2592 compute nodes, 5184 Cavium/Marvell ThunderX2 CPUs, 28 cores/CPU, 2.0 Ghz, 128GB DDR/node
  - Mellanox InfiniBand EDR network
- Importance to ECP and Exascale technology paths
  - Arm provides a diverse alternative to GPU-accelerated nodes through scalable vector extensions and planned integration of on-package high-bandwidth memory
  - Arm business model opens avenues for future HPC-customized on-die and on-package ASICs
  - Innovation-driven leadership for U.S. in a competitive global market
- Status
  - Contract award: May 2018
  - Delivery: September 2018
  - HPL and HPCG benchmarks demonstrated: October 2018
  - Early access codes (MD, DSMC, PIC) have run on over 1/2 of the machine.
  - Moved to classified section: May 2019

Prove-out a new integrator/technology option for future Exascale and capacity HPC systems

JLSE AT ARGONNE

- What is JLSE?
  - The Joint Laboratory for System Evaluation (JLSE) is a collaboration between the Mathematics and Computer Science Division and the Argonne Leadership Computing Facility with the aim of evaluating future high-performance computing platforms.
- HPE Comanche partition
  - 36 compute nodes
  - Dual-socket ARM Marvell ThunderX2 CN9975 processors / node
  - 2.2 GHz reference frequency (2.5 GHz on Turbo mode) with 28 cores/CPU
  - Mellanox InfiniBand EDR network
- Skylake partition
  - 13 compute nodes
  - Dual-socket Intel Xeon Platinum 8180M processors / node
  - 2.5 GHz reference frequency with 28 cores/CPU
  - Intel OPA
- Other partitions
  - KNL partition, Haswell partition, Broadwell partition, V100 partition, Power9 partition, FPGA partition and so on
- https://press3.mcs.anl.gov/jlse/ for more details

CAVIUM* THUNDER X2 SOC



- Now part of Marvel
- Provided by Cavium

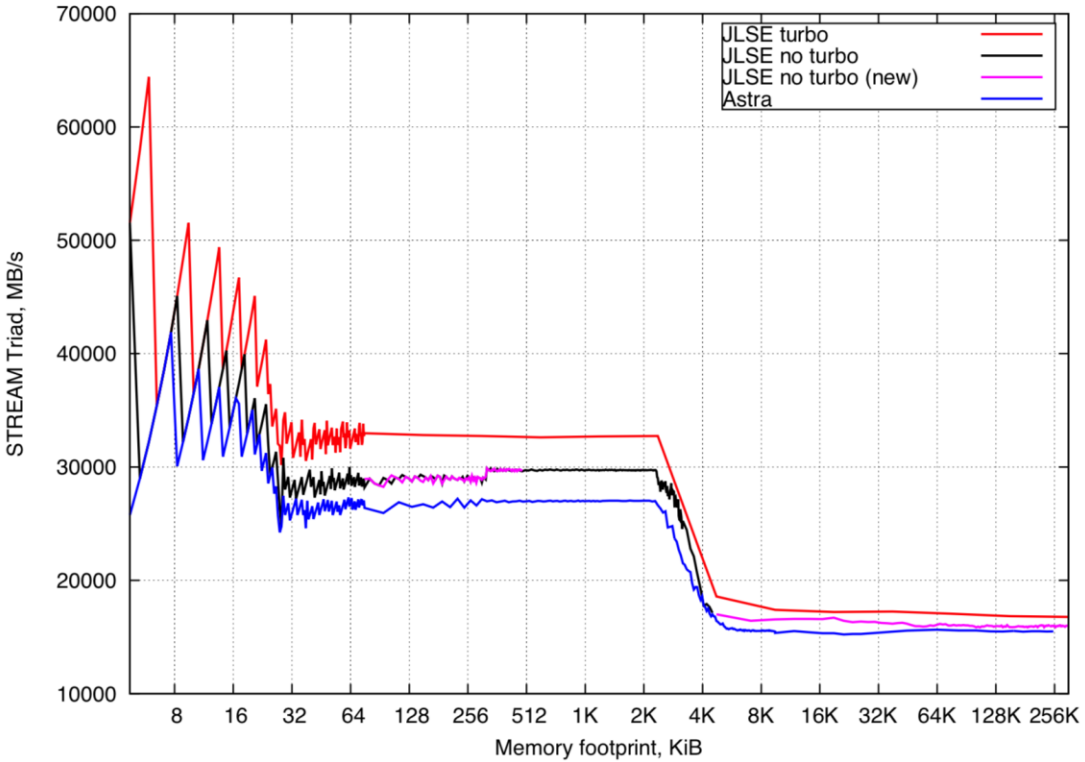
- ❑ 28 core Arm v8.1 2.0 GHz or 2.2 GHz
- ❑ Out of order, SMT1, SMT2, SMT4
- ❑ Dual socket configurations
- ❑ 8 DDR4 memory channels per socket
- ❑ 2 Neon 128-bit wide vector units
- ❑ Up to 4 instructions issued per cycle
- ❑ Arm and GCC compilers
- ❑ ArmPL – BLAS and FFT performance library

Peak numbers

- 8 channels * 2400 MT/s * 8 Bytes/T = 153.6 GB/s
  - Expected 130 GB/s available for applications
- 2 FPUs * 2 way * 2 Flops * 2 GHz = 16 GFlops double precision per core

- Reference KNL based resource
  - About 90 GB/s DDR, about 420 GB/s MCDRAM,
  - About 35.3 GFlops/core
  - 2.2x in Floprate, 0.7x in DRAM BW

STREAM TRIAD SINGLE CORE



32 KiB L1 cache
2 MiB L2 cache bank
    Suspect 8 cores sharing
~17 GB/s/core limit
Turbo when in cache

STREAM TRIAD DUAL SOCKET

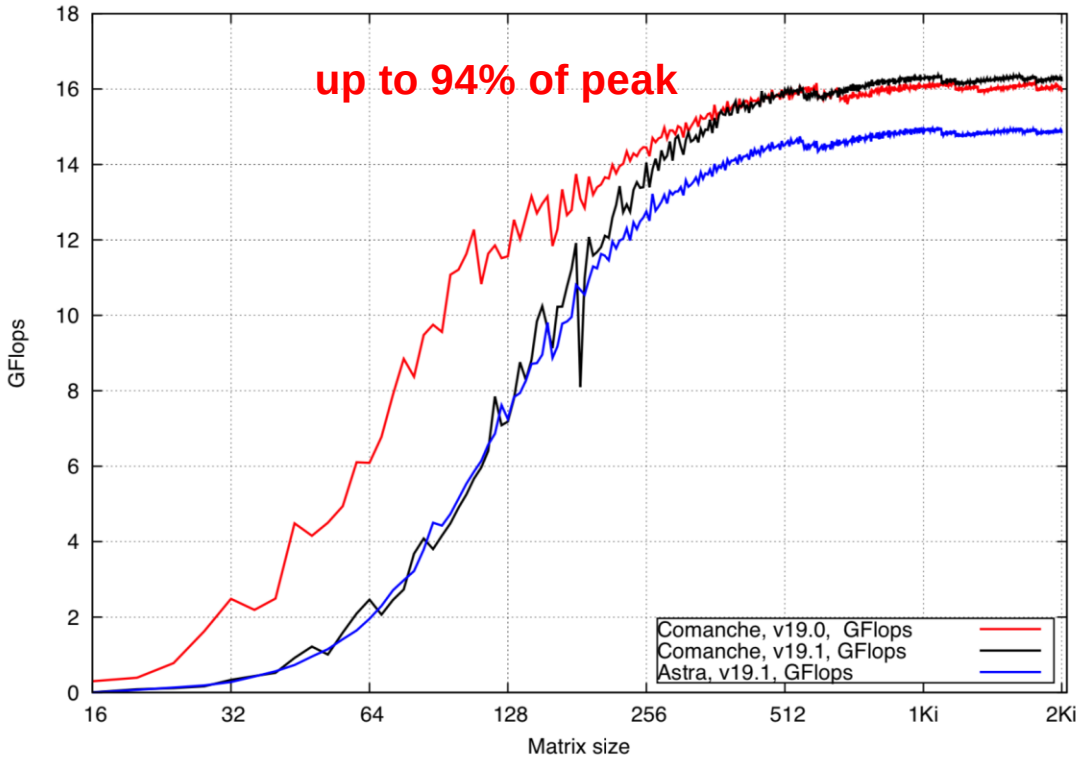
```
armclang -Ofast -mcpu=native -g -fopenmp -ffast-math -ffp-contract=on \
         -DN=7700000000 -c stream.c -o stream.o
armclang -Ofast -mcpu=native -g -fopenmp -ffast-math -ffp-contract=on stream.o -o stream

OMP_NUM_THREADS=112 OMP_PLACES=thread OMP_PROC_BIND=close \
         KMP_AFFINITY=verbose ./stream
```

Function	Rate (MB/s)	Avg time	Min time	Max time
Copy:	222182.1048	0.5603	0.5545	0.5827
Scale:	221797.3972	0.5629	0.5555	0.5848
Add:	225786.0210	0.8224	0.8185	0.8377
Triad:	225792.9929	0.8226	0.8184	0.8320

Target 260 GB/s
Results might be further improved

DGEMM M=N=K SINGLE CORE

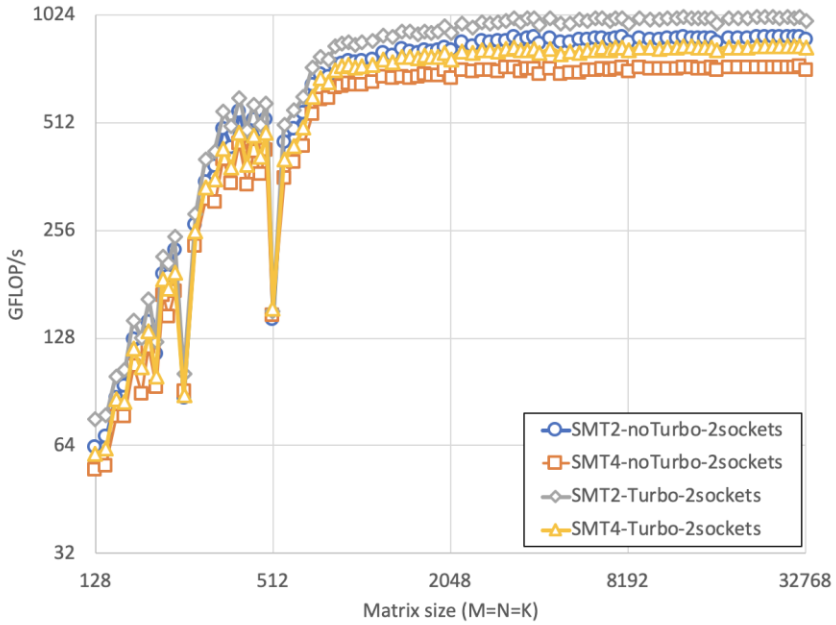


Peak Astra 16 GFlops
Peak Comanche 17.6 GFlops

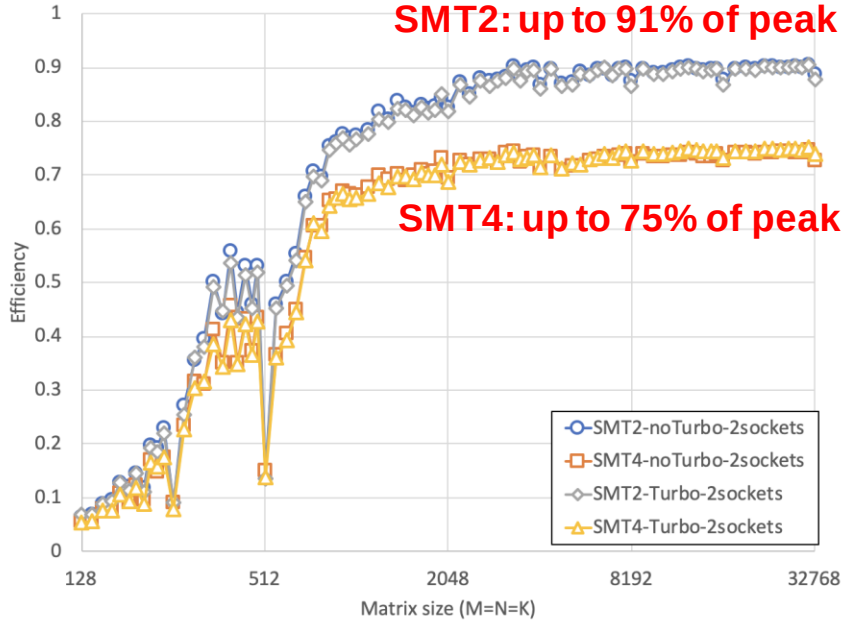
DGEMM M=N=K ON DUAL-SOCKET TX2

SMT2 and SMT4, w/ Turbo and w/o Turbo

DGEMM Flop-rates

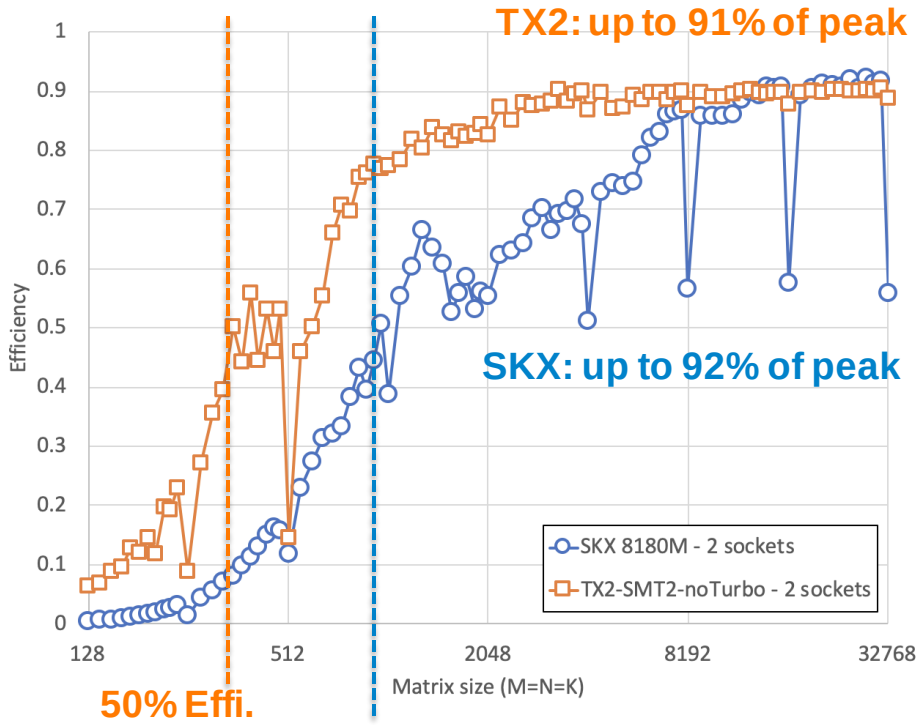


DGEMM Efficiency



DGEMM EFFICIENCY M=N=K ON DUAL-SOCKET

TX2 vs. SKX



- The maximum efficiency
  - TX2: 91%
  - SKX: 92%
  - Both TX2 and SKX shows similar maximum efficiency for large matrices
- Matrix size to achieve 50% efficiency
  - TX2: M=N=K >= 486
  - SKX: M=N=K >= 972
  - TX2 shows higher efficiency than SKX for small matrices

INSTRUCTION THROUGHPUT AND LATENCY

```
L:
add x12, x12, #256
fadd v1.4s, v2.4s, v2.4s
adds x11, x11, #4
b.ne. .L
```

IPC: 3.92

```
L:
fadd v1.4s, v2.4s, v2.4s
fadd v3.4s, v3.4s, v4.4s
adds x11, x11, #4
b.ne. .L
```

IPC: 3.93

fadd, fmul, fmla/fmls, frsqrte, frsqrts, frecpe – 6 cycles

fsqrt – 17 cycles

Suggests
unroll4+SMT2 or
unroll2+SMT4

PERFORMANCE COUNTERS – PAPI 5.6.0+

```
PAPI version             : 5.7.1.0
Operating system         : Linux 4.14.0-
115.5.1.el7a.aarch64
Vendor string and code   : ARM (7, 0x7)
Model string and code    :  (1, 0x1)
CPU revision             : 1.000000
CPUID                    : Family/Model/Stepping
8/175/1, 0x08/0xaf/0x01
CPU Max MHz              : 3
CPU Min MHz              : 3
Total cores              : 112
SMT threads per core     : 2
Cores per socket         : 28
Sockets                  : 2
Cores per NUMA region    : 56
NUMA regions             : 2
Running in a VM          : no
Number Hardware Counters : 6
Max Multiplex Counters   : 384
Fast counter read (rdpmc): no
```

```
PAPI_L1_DCM  0x80000000  Yes   No   Level 1 data cache misses
PAPI_L1_ICM  0x80000001  Yes   No   Level 1 instruction cache misses
PAPI_L2_DCM  0x80000002  Yes   No   Level 2 data cache misses
PAPI_L2_LDM  0x80000019  Yes   No   Level 2 load misses
PAPI_TOT_INS 0x80000032  Yes   No   Instructions completed
PAPI_FP_INS  0x80000034  Yes   No   Floating point instructions
PAPI_LD_INS  0x80000035  Yes   No   Load instructions
PAPI_SR_INS  0x80000036  Yes   No   Store instructions
PAPI_BR_INS  0x80000037  Yes   No   Branch instructions
PAPI_VEC_INS 0x80000038  Yes   No   Vector/SIMD instructions
                                    (could include integer)
PAPI_TOT_CYC 0x8000003b  Yes   No   Total cycles
PAPI_L2_DCH  0x8000003f  Yes   No   Level 2 data cache hits
PAPI_L1_DCA  0x80000040  Yes   Yes  Level 1 data cache accesses
PAPI_L1_DCR  0x80000043  Yes   No   Level 1 data cache reads
PAPI_L2_DCR  0x80000044  Yes   No   Level 2 data cache reads
PAPI_L1_DCW  0x80000046  Yes   No   Level 1 data cache writes
PAPI_L2_DCW  0x80000047  Yes   No   Level 2 data cache writes
PAPI_L1_ICA  0x8000004c  Yes   No   Level 1 instruction cache accesses
```

PAPI_get_real_cyc(), PAPI_get_real_nsec()
– frequency 2.0 GHz, 2.2 GHz, confirmed
PAPI_overflow() – sampling

242 native events reported.

Compatibility constrains

HPM PERFORMANCE TOOL

- Originally developed by IBM's Bob Walkup for Blue Gene/L
  - ✓ Unsupported by IBM, sources available
  - ✓ Native, libc, and PAPI sampling

- Extensively used on all ALCF resources
  - ✓ Blue Gene/Q, KNL, Xeon, Power, and ARM
  - ✓ Text based, small IO requirement, easy to use remotely

- Extensive set of features
  - ➢ Designed by a performance engineer
  - ➢ Shared libraries profiling support
  - ➢ Directed or automatic profiling
  - ➢ Function-level flat profiling
  - ➢ Source file profiling
  - ➢ Statement-level profiling
  - ➢ MPI profiling with communication matrix
  - ➢ Memory footprint measurements
  - ➢ Context switches

HACC: LEADERSHIP COMPUTING

HACC is the only comprehensive cosmology simulation framework capable of running at full scale on ***all*** next-generation supercomputers

- **HACC Physics:** Gravity — Vlasov-Poisson equation solved using a hybrid grid/particle method; Gasdynamics — Euler equation solved using new CRK-SPH algorithm); Subgrid models for gas cooling/heating, star formation and SN/AGN feedback
- **Performance Focus:** Aim for high absolute performance on all HPC platforms, C++/MPI + 'X' programming model
- **Design Features:** Flexible chaining mesh for optimizing short-range solvers, highly optimized force kernels with mixed precision, adaptive time-stepping, task-based load-balancing
- **Analysis:** CosmoTools library (in situ/co-scheduled/offline) for a number of large-scale analysis tasks

- **HACC ASCR Programs**
  - ECP ExaSky & CoPA projects
  - ALCC/INCITE awards
  - NERSC Science at Scale award
  - ESPs (ALCF, NERSC, OLCF)
  - ALCF Data project
  - Aurora Data/Learning project
  - CORAL Benchmarks for next-gen systems at DOE LCFs

HACCMK GRAVITY KERNEL

```
#pragma omp simd
for ( j = 0; j < count1; j++ )
{
    dxc = xx1[j] - xxi;
    dyc = yy1[j] - yyi;
    dzc = zz1[j] - zzi;

    r2 = dxc * dxc + dyc * dyc + dzc * dzc;
    if ( r2 < fsrrmax2 ) {
        f = r2 + mp_rsm2;
        f =  1.0f / ( sqrtf( f * f * f ) ) - ( ma0 + r2*(ma1 + r2*(ma2 + r2*(ma3 + r2*(
             ma4 + r2*ma5))));
        f = mass1[j] * f;

        xi = xi + f * dxc;
        yi = yi + f * dyc;
        zi = zi + f * dzc;
    }
}
```

- ✓ Arm compiler
- ✓ Fully vectorized code

- ❖ fdiv, fsqrt – not frsqrte
- ❖ Dependency chain r2, poly, mass
- ❖ 3 operand FMA semantics
  - a1 = a1 + r2 * a0 vs acc = a1 + r2 * acc
  - 4 operand semantics, scalar only
- ❖ Unroll 2, no choice

HACCMK SINGLE CORE TUNING

	SMT1	SMT2	SMT4	Notes
Time, s	14.56	10.05	**8.47**	Baseline, direct KNL port
IPC	0.7546	1.0962	1.3028	
Time, s	10.28	7.61	6.76	fsqrt , fdiv -> frsqrte
IPC	1.0411	1.4108	1.5932	
Time, s	8.64	5.60	5.94	select, direct fma, registers
IPC	0.938	1.4528	1.3704	
Time, s	5.93	**5.23**		Best known version (**1.62x**)
IPC	1.3667	1.5508		

CORAL HACC GRAVITY-ONLY BENCHMARK

320^3, 5 steps, about 6 GB memory footprint, 16 cores on each system

RPN	OMP	TH	BG/Q Time, s	KNL cache Time, s	KNL flat IPM Time, s	KNL flat DDR Time, s	JLSE Time,s	Astra Time, s
4	4	16	4297	995.0315	984.7330	960.5757	977.8096	1594.0464
4	8	32	2677	825.5221	818.9500	816.7369	**854.3825**	
4	16	64	2504	972.2244	961.7572	956.9851	921.1577	
8	2	16	4362	609.5128	601.9360	600.3612	994.5888	1661.1463
8	4	32	2571	468.6857	463.4524	468.0986	871.0338	
8	8	64	**2278**	437.7195	433.5394	**432.2610**	920.8164	
16	1	16		624.6822	623.1745	624.7912	1219.9419	2099.2842
16	2	32		525.4434	524.6044	524.7912	1072.3818	
16	4	64	2581	478.6204	478.1015	479.3992	1131.2922	

BG/Q: 1 node, 16 cores
KNL: ¼ of the node (quadrant), 16 cores
Thunder, Astra: 1 socket, 16 cores
Xeon Skylake 8176, 8 cores**, 281 seconds**

KNL is used as a baseline

Expected: 2.36x KNL-TX2 core to core

- 4x from vector width, 0.59x from frequency

Achieved: **1.98x**

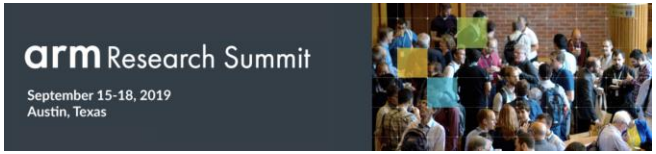
CONCLUDING REMARKS

- **TX2 is a solid matured architecture which can be effectively used for HPC workloads.**
- **TX2 memory subsystem delivers competitive solutions for HPC-centric and general purpose-centric architectures.**
- **We have confirmed the major architectural characteristics such as instruction throughput, floating-point performance, single thread performance, cache, and memory bandwidth.**
- **ARM provides high quality BLAS implementation and compiler technology in their products.**
- **For HPC applications, vector-width in TX2 is insufficient.**
- We presented the results of ALCF application test suite for Intel KNL, Intel Skylake, NVIDIA V100 GPU, and ARM Marvell TX2, and compared their performance from various aspects.

ACKNOWLEDGEMENT

- This Work was supported by the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.
- We also gratefully acknowledge the computing resources provided and operated by the Joint Laboratory for System Evaluation (JLSE) at Argonne National Laboratory.

ARM RESEARCH SUMMIT 2019



Argonne NATIONAL LABORATORY

EARLY EXPERIENCE OF THE ARM-BASED HIGH-PERFORMANCE COMPUTING ECO-SYSTEM

VITALI MOROZOV, JAEHYUK KWACK, THOMAS APPLENCOURT, COLLEEN BERTONI, YASAMAN GHADAR, HUIHUO ZHENG, CHRISTOPHER KNIGHT, AND SCOTT PARKER

Argonne Leadership Computing Facility
Argonne National Laboratory

September 16, 2019, Austin, TX

HPC BENCHMARK/APPLICATION PERFORMANCE

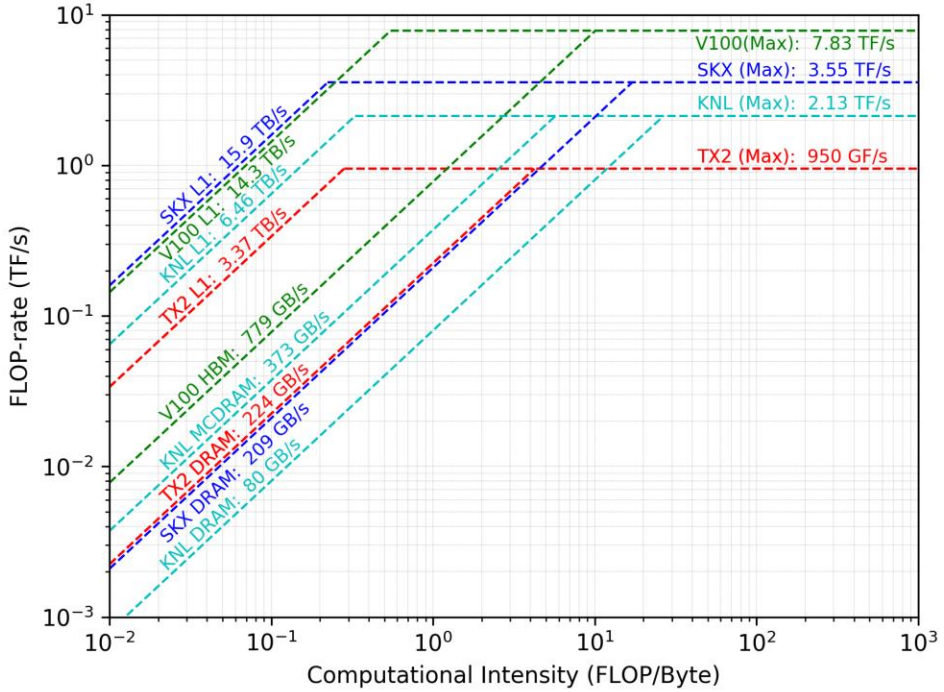
ARM HPC USER GROUP, 2:20 PM TOMORROW

EMPLOYED PROCESSORS

- Processor models
  - KNL: a single Intel KNL 7320 processor
  - TX2: dual-socket Marvell ThunderX2 CN9975 processors
  - SKX: dual-socket Intel Xeon Platinum 8180M processors (Skylake)
  - V100: a single NVIDIA V100-SMX2 GPU

- Measured Peak performance via ERT (LBL.gov)

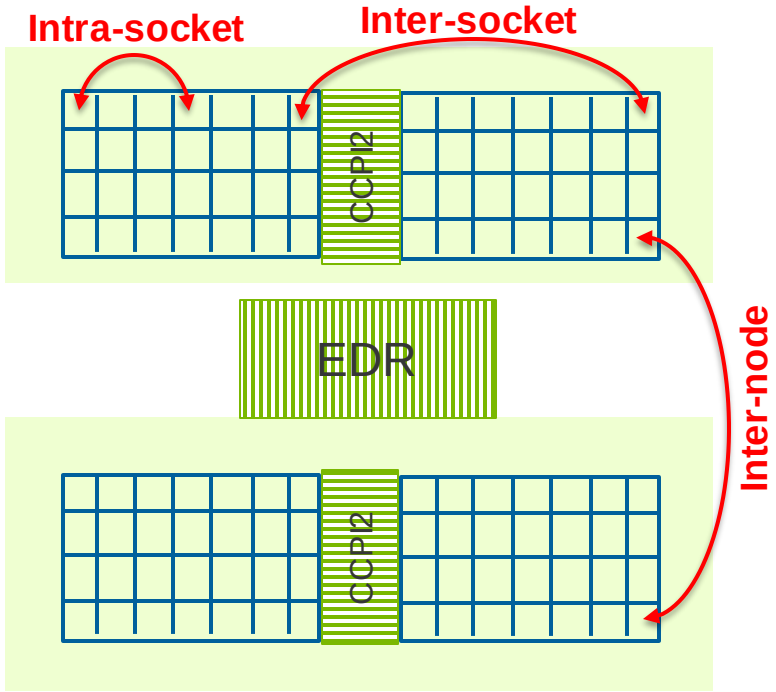
	Flop-rate (TF/s)	L1 (TB/s)	L2 (TB/s)	LLC (GB/s)	DRAM (GB/s)
KNL	2.13	6.46	1.91	373	80(2)
TX2	0.95	3.37	2.63	1091	224
SKX	3.55	15.91	4.55		209
V100	7.83	14.34(1)	3.35		779



(1)Theoretical peak, (2)Stream Triad

MPI MESSAGING-RATES AND LATENCY

Intra-socket vs. Inter socket vs. Inter-node



- MPI messaging-rate with 0 byte
  - Intra-socket: 3.35 mmps (Million-Message/sec)
  - Inter-socket: 2.24 mmps
  - Inter-node: 0.75 mmps - 0.66 mmps
- MPI Latency (= 1/mmps)
  - Intra-socket: around 300 ns
  - Inter-socket: around 450 ns
  - Inter-node: around 1.3 µs – 1.5 µs

ARGONNE TEST SUITE

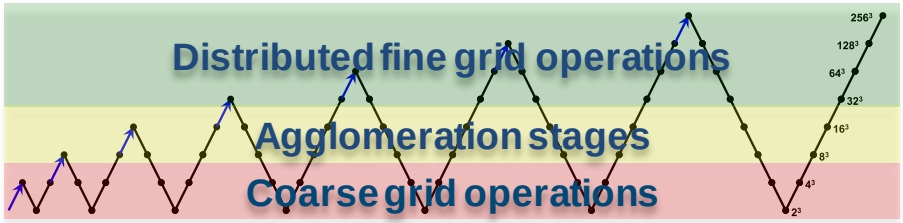
- HPGMG-FV: an ECP proxy application
- NEKBONE: an ECP proxy application and DOE CORAL-2 benchmark
- GAMESS: an ECP application
- LAMMPS: an ECP application and DOE CORAL-2 benchmark
- QMCPACK: an ECP application and DOE CORAL-2 benchmark
- QBOX: an ECP application

  - DOE: U.S. Department of Energy
  - ECP: Exascale Computing Project
  - CORAL: Collaboration of Oak Ridge, Argonne, and Livermore

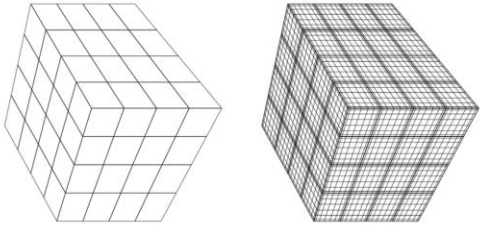
PoC: JaeHyuk Kwack

HPGMG

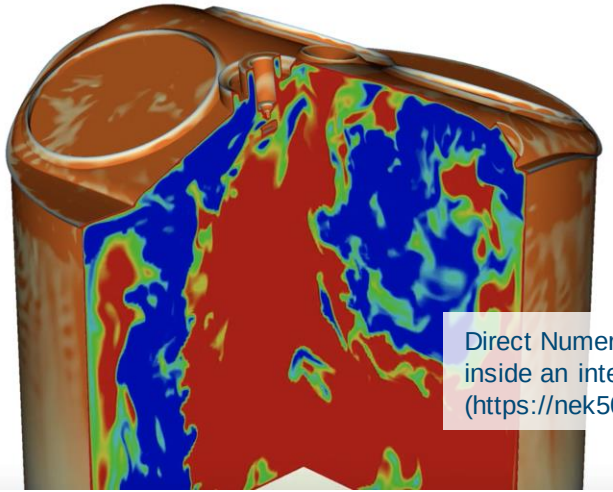
- HPGMG-FE (Finite Element): compute-intensive and cache-intensive
- HPGMG-FV (Finite Volume): memory bandwidth-intensive
  - Used for the list of ranking supercomputers
  - Solving an elliptic problem on isotropic Cartesian grids with 4th order accuracy
  - 4× FP ops, 3× MPI messages, 2× MPI message size w/o DRAM data movement compared to 2th order HPGMG-FV
  - Employing the Full Multi-grid (FMG) F-cycle
  - A series of progressively deeper geometric multi-grid V-cycles



NEKBONE



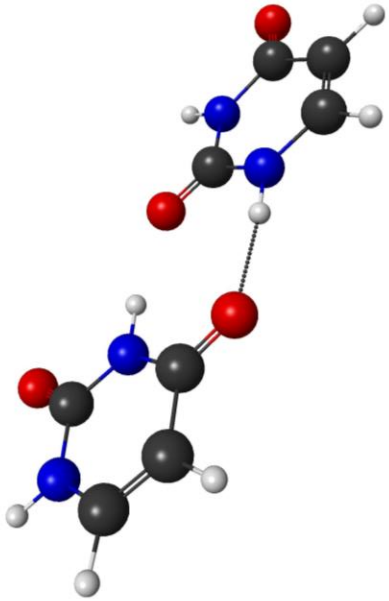
- A mini-app derived from the Nek5000 CFD code which is a high order, incompressible Navier-Stokes CFD solver based on the spectral element method.
- Standard Poisson equation in a 3D box domain with a block spatial domain decomposition among MPI ranks.
- Solver phase: conjugate gradient iterations in an element-by-element fashion.
  - Vector operations
  - Matrix-matrix multiply operations
  - Nearest-neighbor communication
  - MPI Allreduce operations.
- Source:
  - written in C and Fortran
  - MPI+OpenMP



Direct Numerical Simulation of the flow inside an internal combustion engine (https://nek5000.mcs.anl.gov)

GAMESS

- A general quantum chemistry and *ab initio* electronic structure code.
  - *ab initio* SCF energies (e.g. RHF and MCSCF)
  - Force fields (e.g., the Effective Fragment Potential)
  - Perturbative corrections to Hartree-Fock (e.g., MP2 and RI-MP2)
  - Near-linear scaling fragmentation methods (e.g., Fragment Molecular Orbital method)
  - *ab initio* gradients, hessians, and geometry optimizations.
- Source
  - Mainly written in Fortran
  - An MPI parallelization library (DDI library) written in C
  - An optional C++ library with re-implementations of certain methods
  - MPI + X
    - OpenMP for CPU cores
    - CUDA for NVIDIA GPU accelerators.



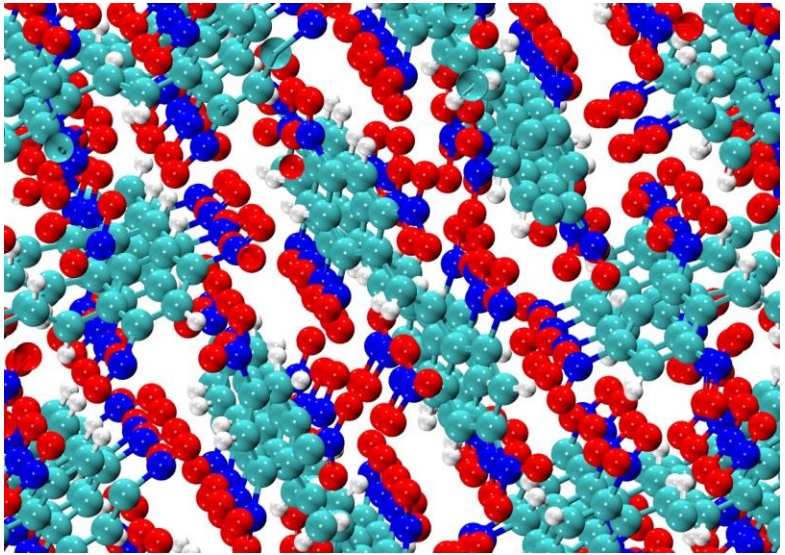
Generated by wxMacMolPlt

PoC: Yasaman Ghadar, Christopher Knight

LAMMPS

- A classical molecular simulation code commonly used for modeling various states of matter, liquids, surfaces, solids, and biopolymers. It supports multiple physical models, particle types, and sampling methods.

- Source
  - Written in C/C++
  - Parallelized with MPI + X
    - X for OpenMP, CUDA, OpenCL, Kokkos, and explicit vectorization
  - An unmodified version of LAMMPS, 19Feb19
  - DOE CORAL-2 LAMMPS benchmark
    - Analysis of the reactive forcefield ReaxFF



QMCPACK

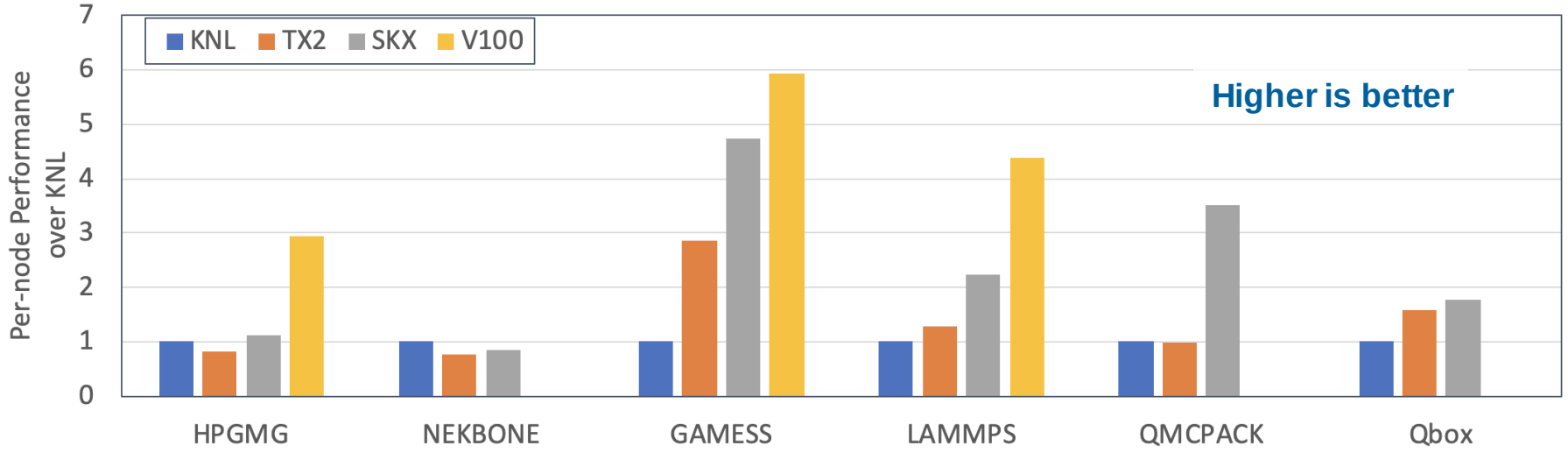
- An open source quantum Monte Carlo package for *ab-initio* electronic structure calculations.
- It supports calculations of metallic and insulating solids.
- It uses a Metropolis Monte Carlo algorithm which generates samples sequentially via a random walk along a Markov chain.
- Each OpenMP thread executes an independent Markov chain or a walker. After each walker has completed a number step, the simulation is completed. Hence, the more workers you have, faster to the method converges.
- Our figure of merit (FOM) measures how many walkers have been moved in one second.
- Version: QMCPACK v3.7.0
- Input (a.k.a. S32)
  - 32 repeats of a NiO primitive cell leading to 128 atoms and 1536 electrons

QBOX

- A C++ MPI/OpenMP scalable parallel implementation of first-principles molecular dynamics based on the plane-wave, pseudopotential density functional theory
- It uses FFTW for 3D Fast Fourier Transformation and ScaLAPACK for parallel dense linear algebra.
- Linking against the vendor provided libraries
  - MKL on SKX and KNL
  - ArmPL on TX2
- Input
  - A silicon carbide periodic solid system which contains 64 atoms (32 silicon and 32 carbon atoms) and 256 electrons
  - Performing the ground state calculation using PBE0 hybrid functional
  - Total number of self-consistent iterations set 5
- Runtime environments
  - 1 MPI process per core on all architectures
  - MPI processes are arranged in a two dimensional array (8 × 7 on SKX/TX2, 8 × 8 on KNL).

PER-NODE PERFORMANCE

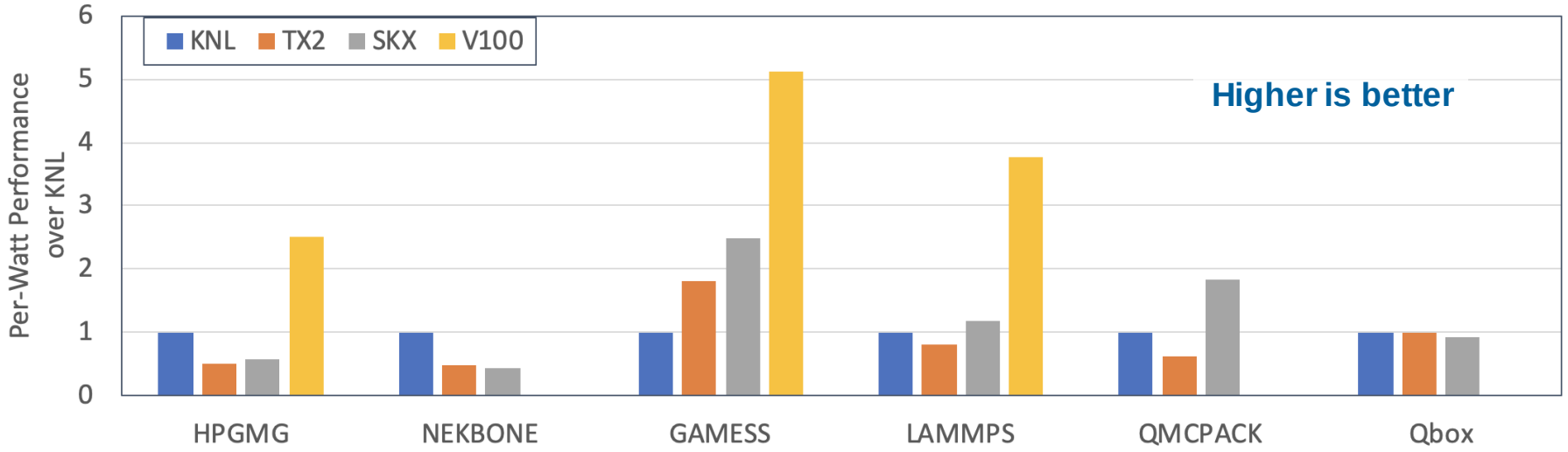
Performance on the same number of nodes



PER-WATT PERFORMANCE

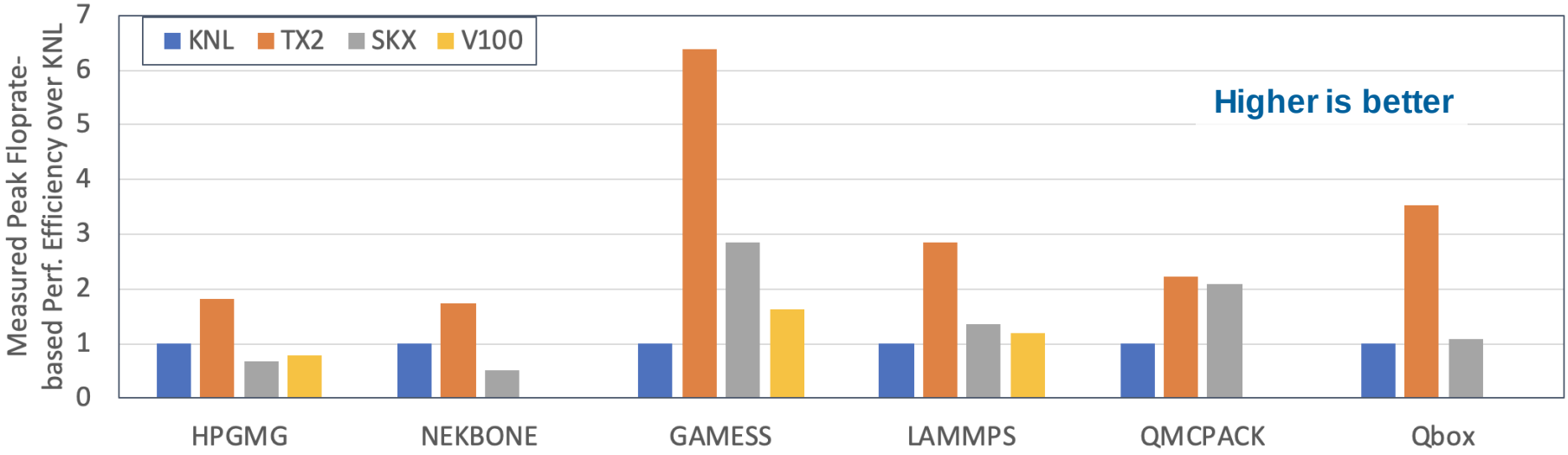
Performance with the same power usage

- TDP (Thermal Design Power)
  - KNL: 215W/socket, 215W/node
  - TX2: 170W/socket, 340W/node
  - SKX: 205W/socket, 410W/node
  - V100: 250W/socket



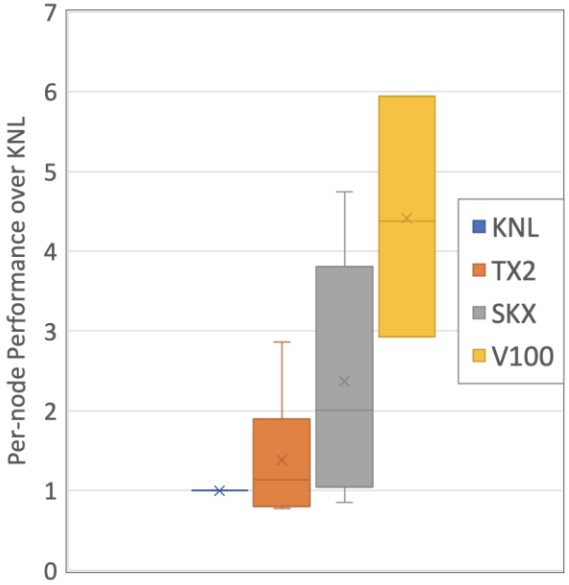
PEAK FLOPRATE-BASED PERF. EFFICIENCY

Performance efficiency on systems with the same peak (e.g., on 2PF systems based on different processors)

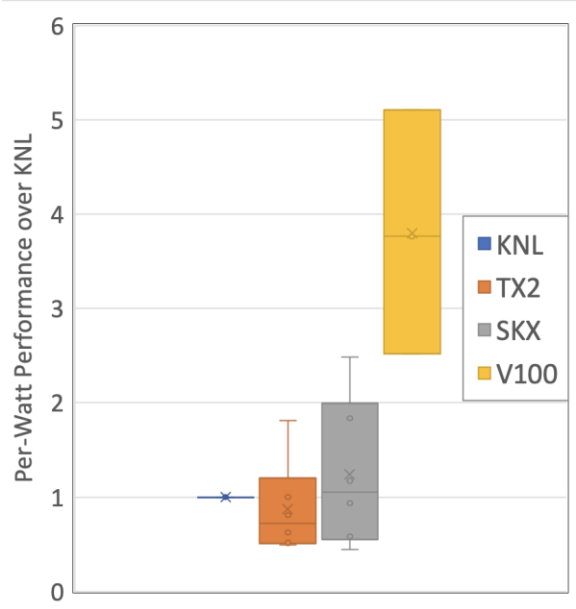


SUMMARY

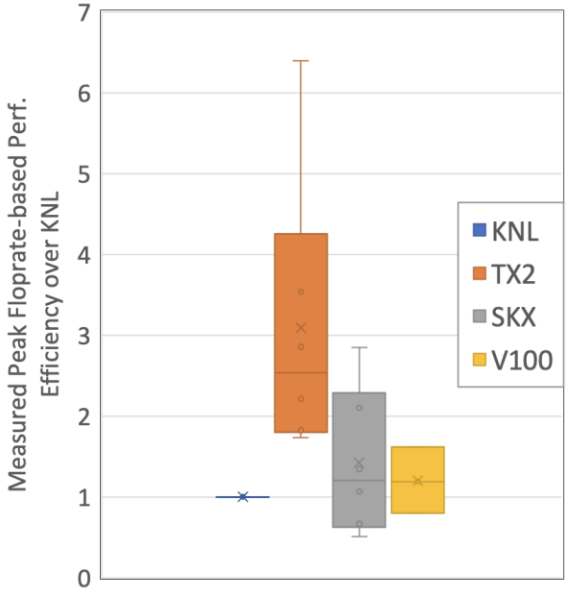
On the same number of nodes



With the same power usage



On systems with the same peak



CONCLUDING REMARKS

CONCLUDING REMARKS

- TX2 is a solid matured architecture which can be effectively used for HPC workloads.
- TX2 memory subsystem delivers competitive solutions for HPC-centric and general purpose-centric architectures.
- We have confirmed the major architectural characteristics such as instruction throughput, floating-point performance, single thread performance, cache, and memory bandwidth.
- ARM provides high quality BLAS implementation and compiler technology in their products.
- For HPC applications, vector-width in TX2 is insufficient.
- We presented the results of ALCF application test suite for Intel KNL, Intel Skylake, NVIDIA V100 GPU, and ARM Marvell TX2, and compared their performance from various aspects.

ACKNOWLEDGEMENT

- This Work was supported by the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.
- We also gratefully acknowledge the computing resources provided and operated by the Joint Laboratory for System Evaluation (JLSE) at Argonne National Laboratory.

THANK YOU!

BACKUP SLIDES