EARLY EXPERIENCE OF
THE ARM-BASED HIGH-PERFORMANCE
COMPUTING ECO-SYSTEM
W/ HPC BENCHMARKS AND APPLICATIONS

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

Argonne Leadership Computing Facility
Argonne National Laboratory

September 17, 2019, Austin, TX
EMPLOYED ARCHITECTURES & THEIR CHARACTERISTICS
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

- 28 core Arm v8.1 @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 * 2.2 GHz = 17.6 GFlops double precision per core

Reference KNL based resource
- About 90 GB/s DDR, about 420 GB/s MCDRAM,
- About 35.3 GFlops/core

*(Provided by Cavium)

*Now part of Marvell*
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)

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

(1) Theoretical peak, (2) Stream Triad
DGEMM $M=N=K$ ON DUAL-SOCKET TX2
SMT2 and SMT4, w/ Turbo and w/o Turbo

DGEMM Flop-rates

DGEMM Efficiency

SMT2: up to 91% of peak
SMT4: up to 75% of peak
DGEMM EFFICIENCY M=N=K

TX2 vs. SKX vs. V100

- The maximum efficiency
  - TX2: 91%
  - SKX: 92%
  - V100: 87%
  - TX2 shows competitive efficiency for large matrices

- Matrix size to achieve 50% efficiency
  - TX2: M=N=K >= 486
  - SKX: M=N=K >= 972
  - V100: M=N=K >= 1420
  - TX2 shows higher efficiency than SKX and V100 for small matrices
MPI MESSAGING-RATES AND LATENCY

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

- MPI messaging-rate with 0 byte
  - Intra-socket: 3.35 mmmps (Million-Message/sec)
  - Inter-socket: 2.24 mmmps
  - Inter-node: 0.75 mmmps - 0.66 mmmps

- MPI Latency ( = 1/mmmps)
  - Intra-socket: around 300 ns
  - Inter-socket: around 450 ns
  - Inter-node: around 1.3 µs – 1.5 µs
HPC BENCHMARK/APPLICATION PERFORMANCE
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
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 2nd order HPGMG-FV
  - Employing the Full Multi-grid (FMG) F-cycle
  - A series of progressively deeper geometric multi-grid V-cycles
NEKBOiNE

- 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.
  - Vector operations
  - Matrix-matrix multiply operations
  - Nearest-neighbor communication
  - MPI Allreduce operations.
- Source:
  - written in C and Fortran
  - MPI+OpenMP

PoC: Scott Parker

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.
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

Higher is better
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

Higher is better
PEAK FLOPRATE-BASED PERF. EFFICIENCY

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

Higher is better
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!