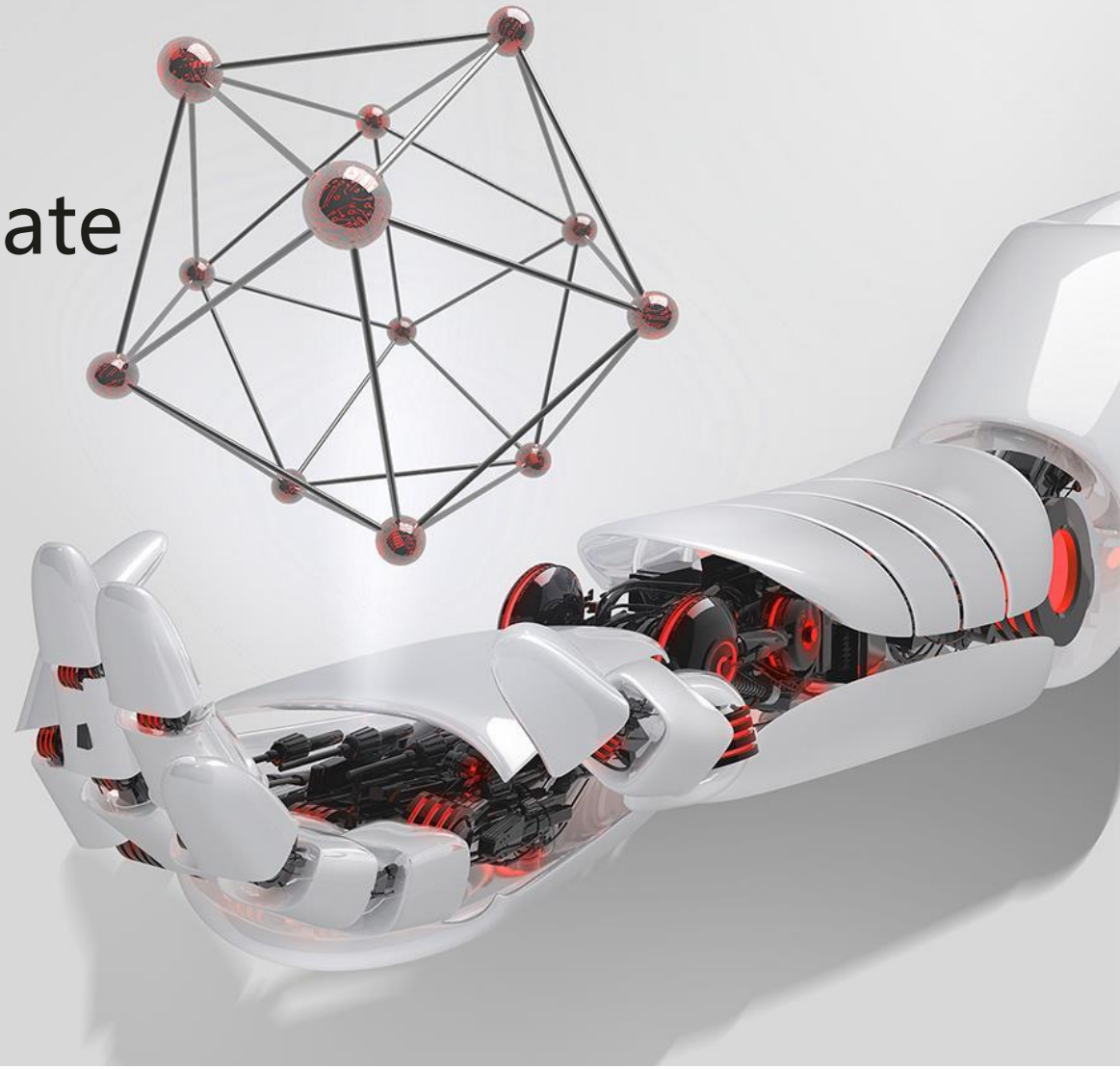
Huawei ARM HPC Software Update

The 2nd R-CCS International Symposium

Zhaohui Ding (丁肇辉)
dingzhaohui@huawei.com
Director, HPC Lab,
Cloud & AI Business Group, Huawei Technologies

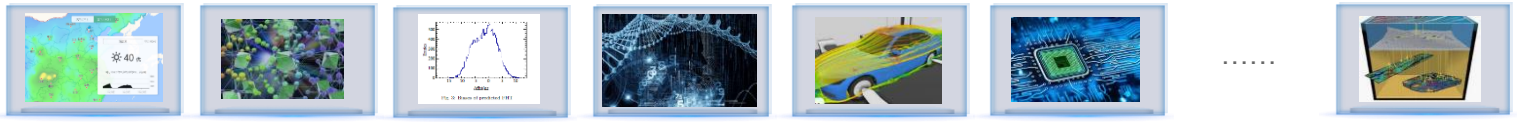


HUAWEI

Agenda

HUAWEI

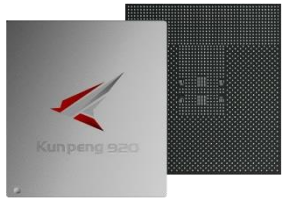
Huawei Kunpeng HPC Solution Stack



Meteorology | Computational Chemistry | Physics | Life Science | CAE | EDA | …… | Reservoir Simulation

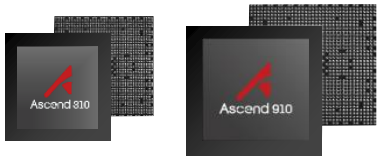
Workload Manager | Application Portal/Workflow | Cluster Management | Monitoring & Dashboard

Huawei Cloud Compiler | Huawei MPI | Math Libraries | Runtime | Tracer, Profiler, Debugger, Porting Advisor



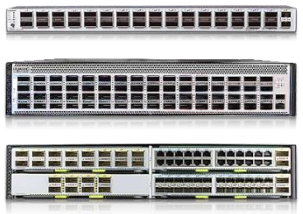
Kunpeng 920

Kunpeng Mainboard



Ascend 310 & 910

FusionStorage Distributed Storage



Ethernet Switches

HUAWEI

The Roadmap of Kunpeng

Hi1612

32 cores@16FF
2.1 GHz
4*64bit DDR3/4
PCIe 3.0 /SAS3.0/10GE

Kunpeng 916

32 cores@16FF+
2.4 GHz
4P
4*64bit DDR3/4
PCIe 3.0 /SAS3.0/10GE

Kunpeng 920

Up to 64 cores@7nm
Up to 3.0 GHz
4P
8 DDR4 Channels
CCIX
RoCE v2
PCIe 4.0 /100GE

Kunpeng 930

Kunpeng 950

Hi1610

16 cores@16FF
2.1 GHz
2*64bit DDR3/4
PCIe 3.0 /SAS3.0/10GE

Available 920 models in 2019 Jun:

- Kunpeng 920-6430: 64C@3.0GHz, 200W
- Kunpeng 920-6426: 64C@2.6GHz, 180W
- Kunpeng 920-4826: 48C@2.6GHz, 150W
- Kunpeng 920-3226: 32C@2.6GHz, 120W

Production
Development
Planning

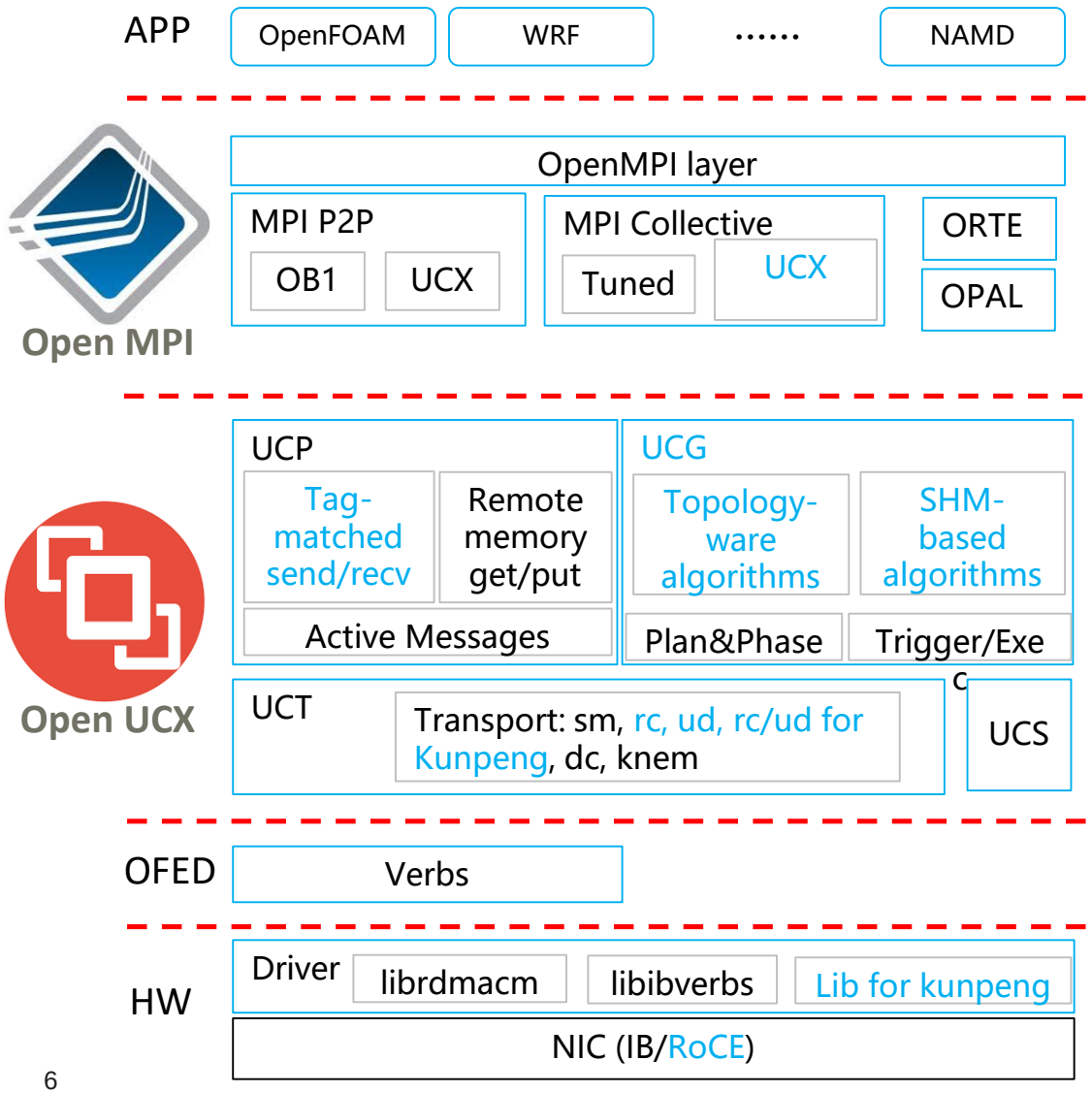
2014 **2016** **2018** **2020** **2022**

HUAWEI

Agenda

HUAWEI

Huawei MPI



Open MPI & Open UCX

- Open MPI is modular, and easily extendable MPI implementation
- Open UCX is communication framework created by industry, laboratories and academia.
- OpenMPI's p2p module is using Open UCX

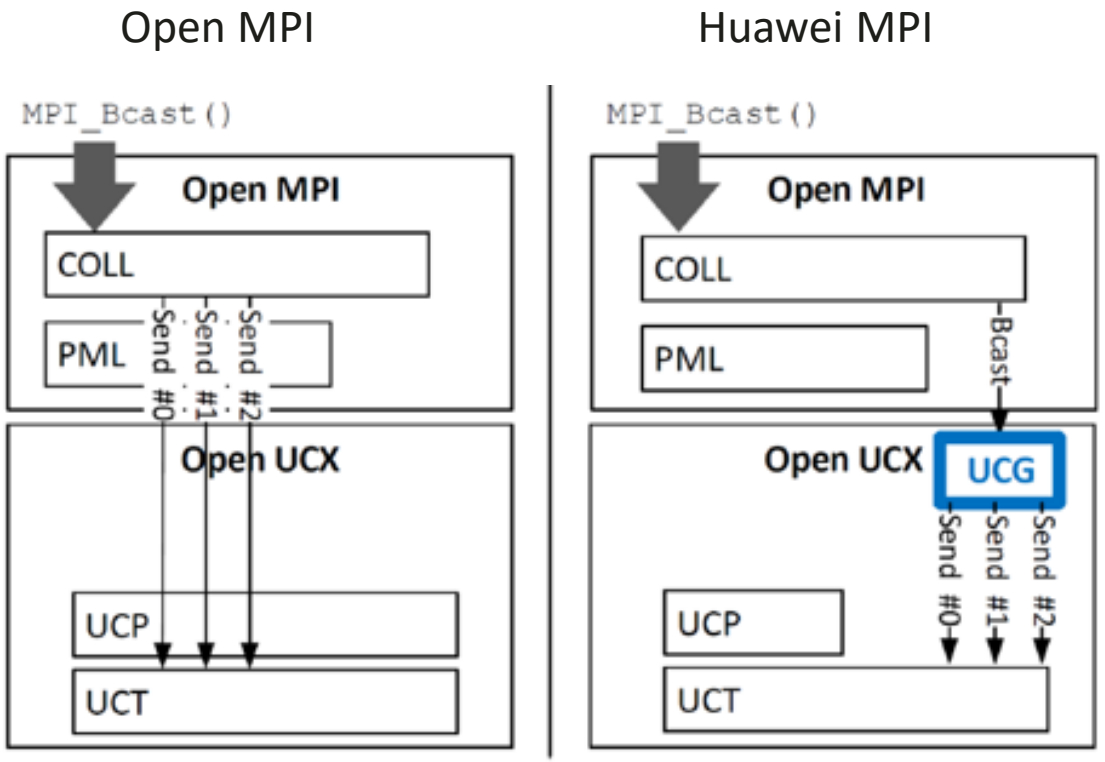
Huawei MPI

- An optimized implementation based on Open MPI
- Proposed UCG(Groups) – collective operations API to UCX
- Implement optimized collective operations algorithms based on Open UCX
- Support ARM and x86
- Optimized topology-aware, SHM algorithms for Kunpeng CPU
- Communication offloading for p2p and coll
- Computing offload in switch

HUAWEI

Consolidating UCX P2P Communication

- Open UCX, as most P2P libraries, contains considerable logic on how best to send data.
- During collective operations, messages of the same size are often sent consecutively. This makes most of the P2P logic redundant.
- Proposed UCG(Groups), move the collective operations logic down to UCX layer
- Consolidating the per-message logic, making it per-collective, to save latency
- Decoupling planning and execution, the collective algorithms can be extended easily by 3rd party.

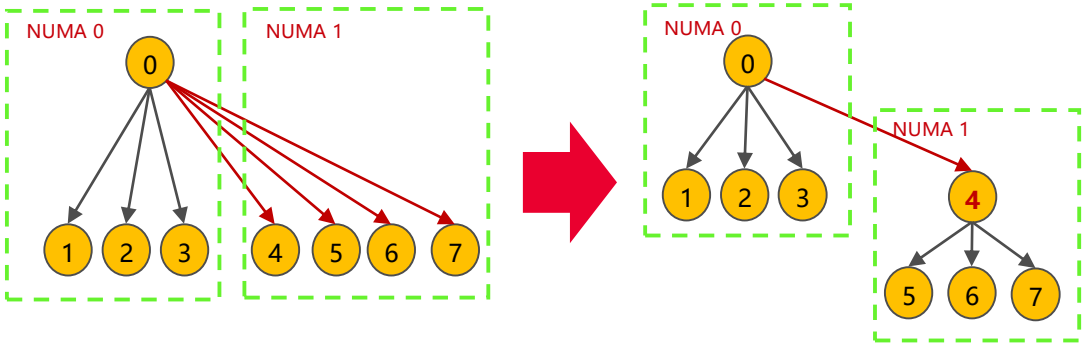


"RDMA-Based Library for Collective Operations in MPI", Alex Margolin and Amnon Barak, ExaMPI'19

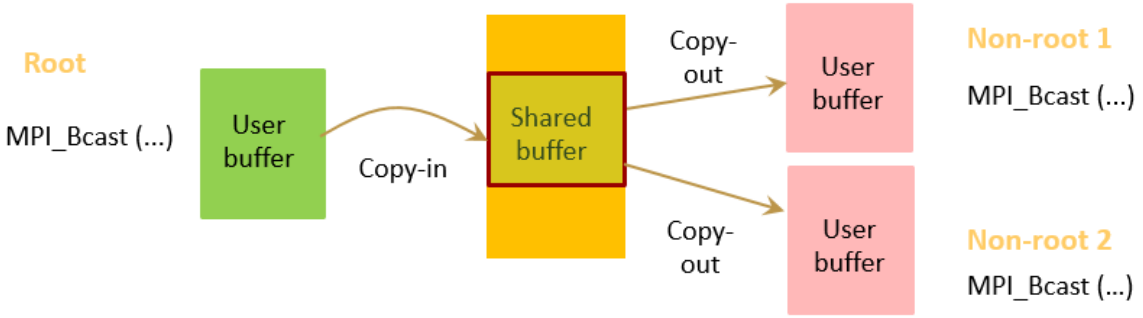
HUAWEI

Collective Algorithms

- Optimized MPI_bcast, MPI_allreduce, MPI_barrier
- NUMA aware and SHM-based intra-node collectives
- Collective operation algorithms
  - Bcast: Binomial, Knomial
  - Allreduce: Recursive doubling, Binomial, Knomial, Ring
  - Barrier: Recursive doubling, Binomial, Knomial
- All the algorithms are implemented under UCG, binomial Bcast, Recursive doubling Allreduce and Barrier are open source.



NUMA-awared MPI_bcast

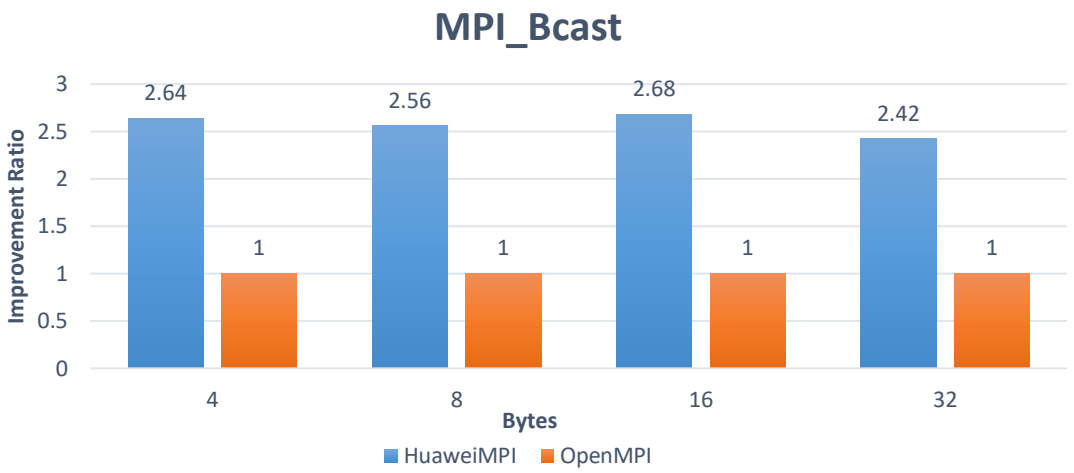


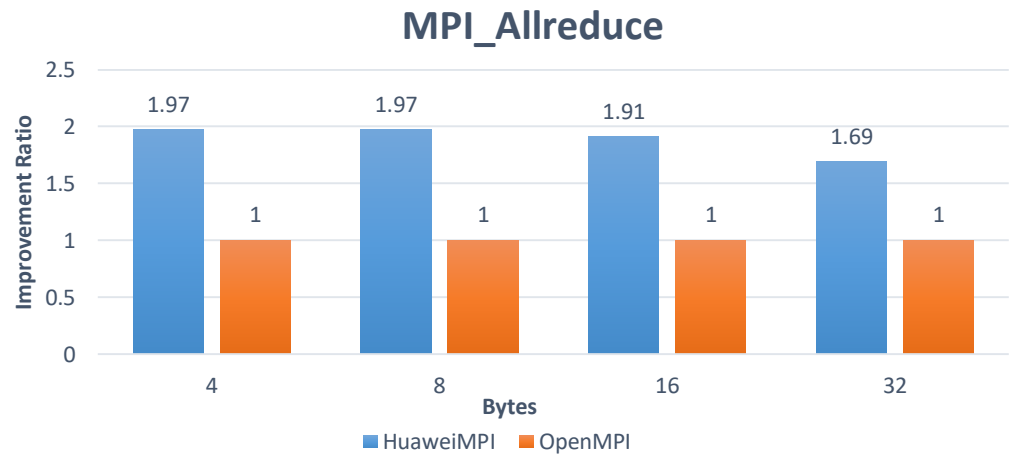
SHM-based MPI_bcast

HUAWEI

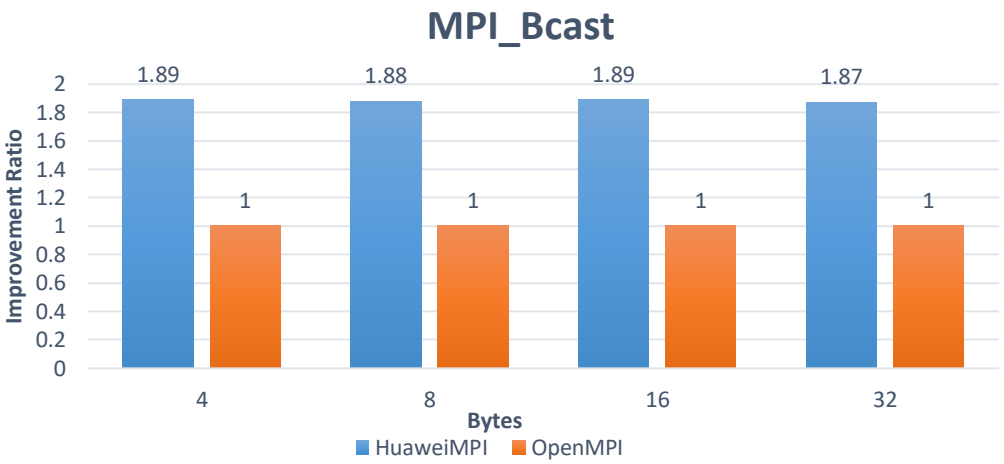
8 Nodes Small Package Benchmark

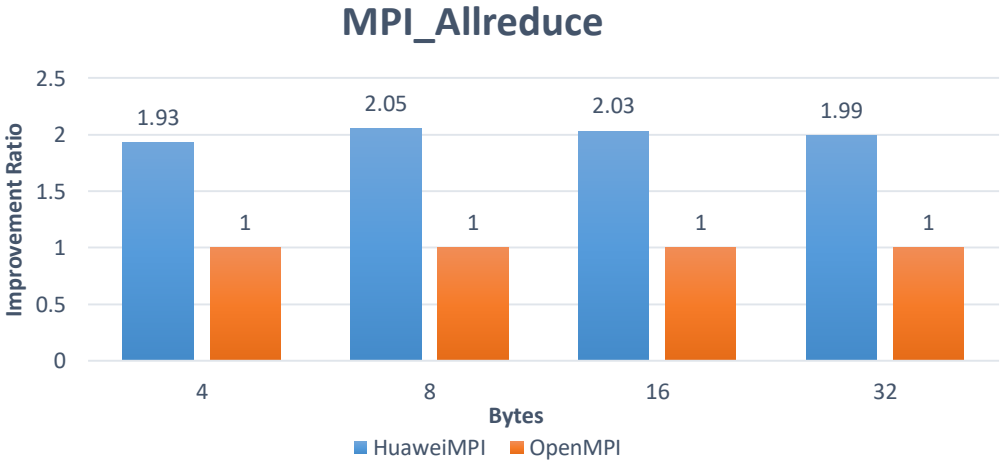
40ppn, IB, Intel Skylake 6148





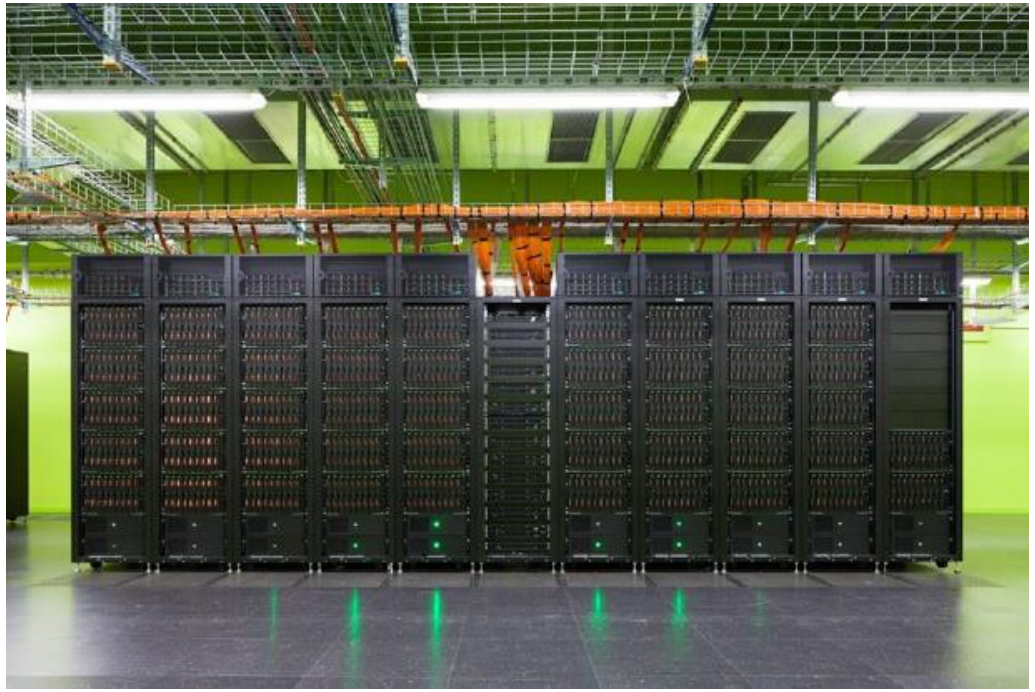
96ppn, IB, KunPeng920





HUAWEI

256 Nodes Benchmark



The table shows the latency (microseconds) of Allreducewith increasing message sizes (Bytes), on 256 nodes (*Intel Xeon E5-2680*CPUs @2.5GHz, Connect-X3 NIC).

Length (bytes)	**Huawei MPI** - Inline Sends		**Huawei MPI** - Buffer Copy		**Open MPI** Latency (us)
	Latency (us)	Improvement	Latency (us)	Improvement	
16	19.5	6.7%			20.9
32	22.78	1.3%			23.08
64	21.57	8.5%			23.58
128	25.79	8.1%			28.07
256			24.95	6.6%	26.7
512			28	9.9%	31.06
1024			29.66	16.0%	35.33
2048			36.03	8.7%	39.47
4096			47.21	6.4%	50.43
8192			66.68	8.7%	73.02
16384			89.07	90.4%	923.65
32768			135.4	88.2%	1142.66
65536			227.83	81.2%	1210.49
131072			442.31	66.1%	1302.86
262144			905.43	39.8%	1503.8
524288			1823.7	2.9%	1878.85

HUAWEI

Agenda

HUAWEI

HCC – Huawei Cloud Compiler

- Huawei Compiler Lab, 300+ people, multiple GCC and LLVM committers

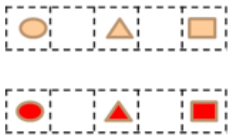
- Develop various compilers for mobile, IoT, network devices and Kunpeng
- HCC for HPC 1.0
  - Instrument Pipeline optimization for Kunpeng 920
  - Optimized mathlib
  - A series of features and optimizations for Kunpeng
- Future plan
  - Based on latest gcc
  - Optimized OpenMP runtime
  - More auto parallelization
  - llvm-based compiler

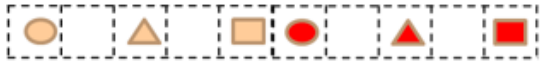
HCC – Optimized Instrument Pipeline

Code:

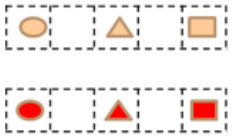
V1 = V2 + V3	K1 = K2 + K3
V0 = V1 - V2	K0 = K1 - K2
V5 = V0 * V2	K5 = K0 * K2

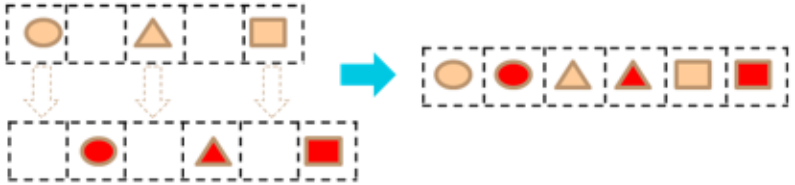
No Optimization



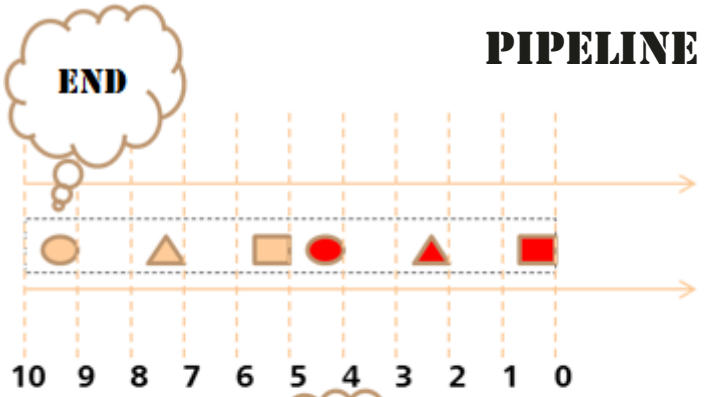


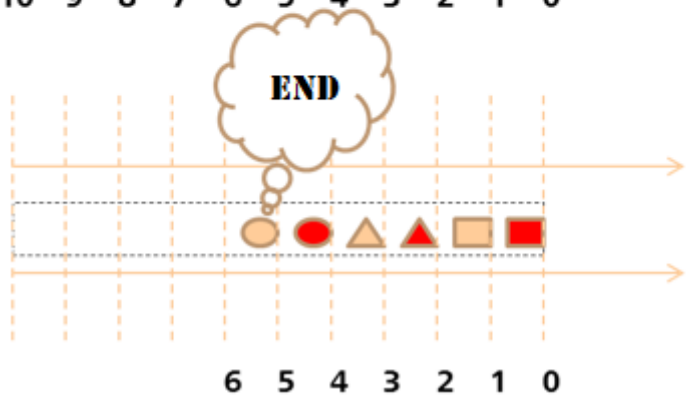
Pipeline Optimization





PIPELINE

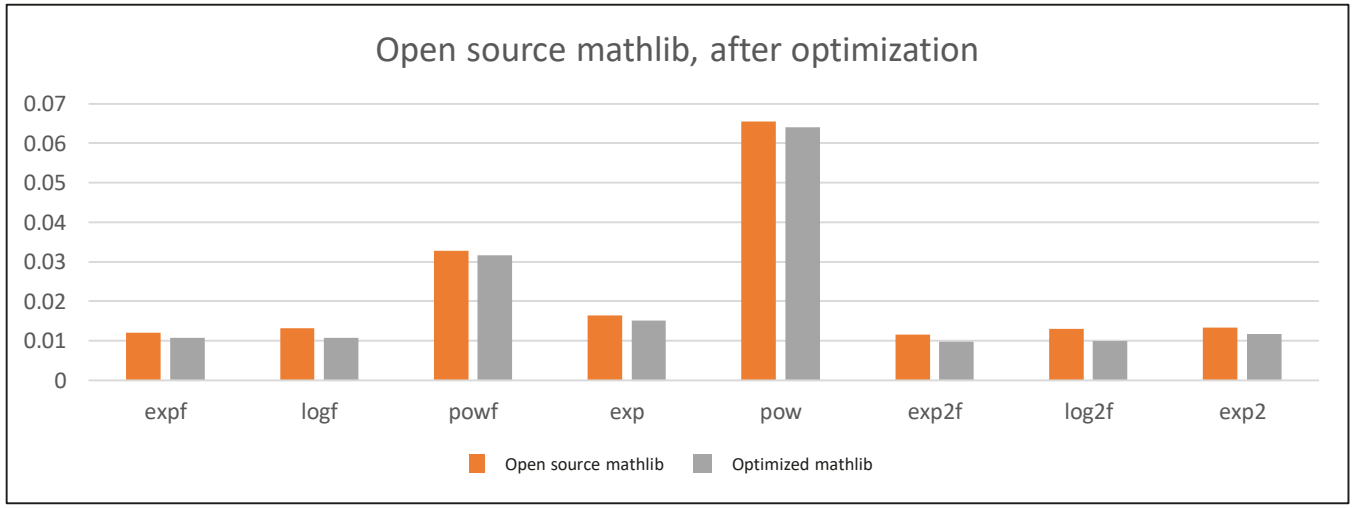




HUAWEI

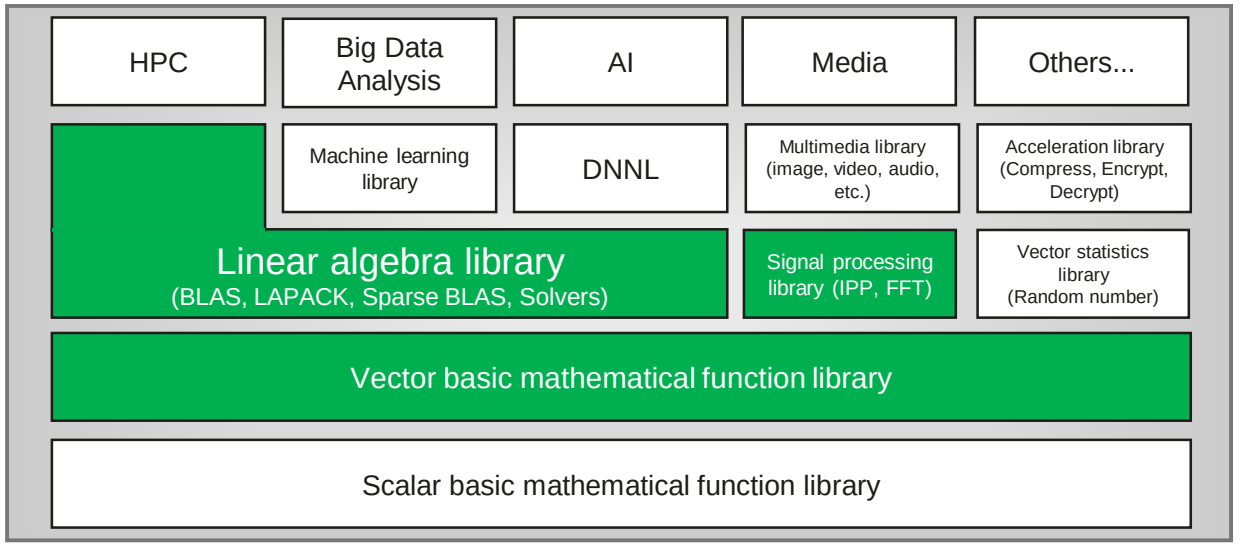
HCC – Feature Highlights

- -fhcc-gfortran-minmax
  - Uses cpu instruments to do the min/max comparison
- -mcmodel
  - Support tiny, small, medium and large
- -DAARCH64_QUADMATH
  - Support 128 bits floating
- - lstringlib -Wl,--wrap=memset
  - memset optimization
- -lmathlib
  - An optimized math library



MAL – Math Acceleration Library

- MAL is Huawei internal project launched from 2016, to provide high performance math libraries for ARM
- MAL includes BLAS, FFT, Lapack, Sparse BLAS
- Open source community also optimize code for Kunpeng, e.g., OpenBLAS



Agenda

HUAWEI

The Challenges of Modern HPC Scheduler

- The new challenges brought by the convergence of HPC, HPDA and Deep Learning
  - More kinds of application frameworks run in same environments
    - The traditional HPC: MPI + X
    - The Bigdata: Hadoop, Spark, Flink etc.
    - The Deep Learning: Tensorflow, PyTorch, paddle etc.
  - How to run all the workload in one HPC cluster? Without changing user habits of the frameworks.

- More and more the task-based applications
  - Tasks scheduling handled by the framework: the challenge is elastic allocation
  - Task scheduling handled by the scheduler: the challenge is job throughput
  - Kunpeng920 has 128 cores, 8k nodes means 1M cores

- More new components than CPU and Job
  - Accelerator, Container job, Resource Bursting, I/O & Storage
  - Take GPU as example, it is not only a “number”, but also should be applied in the policies, fairshare/threshold/reservation/preemption, and in the monitoring & reporting

Donau Design Principle

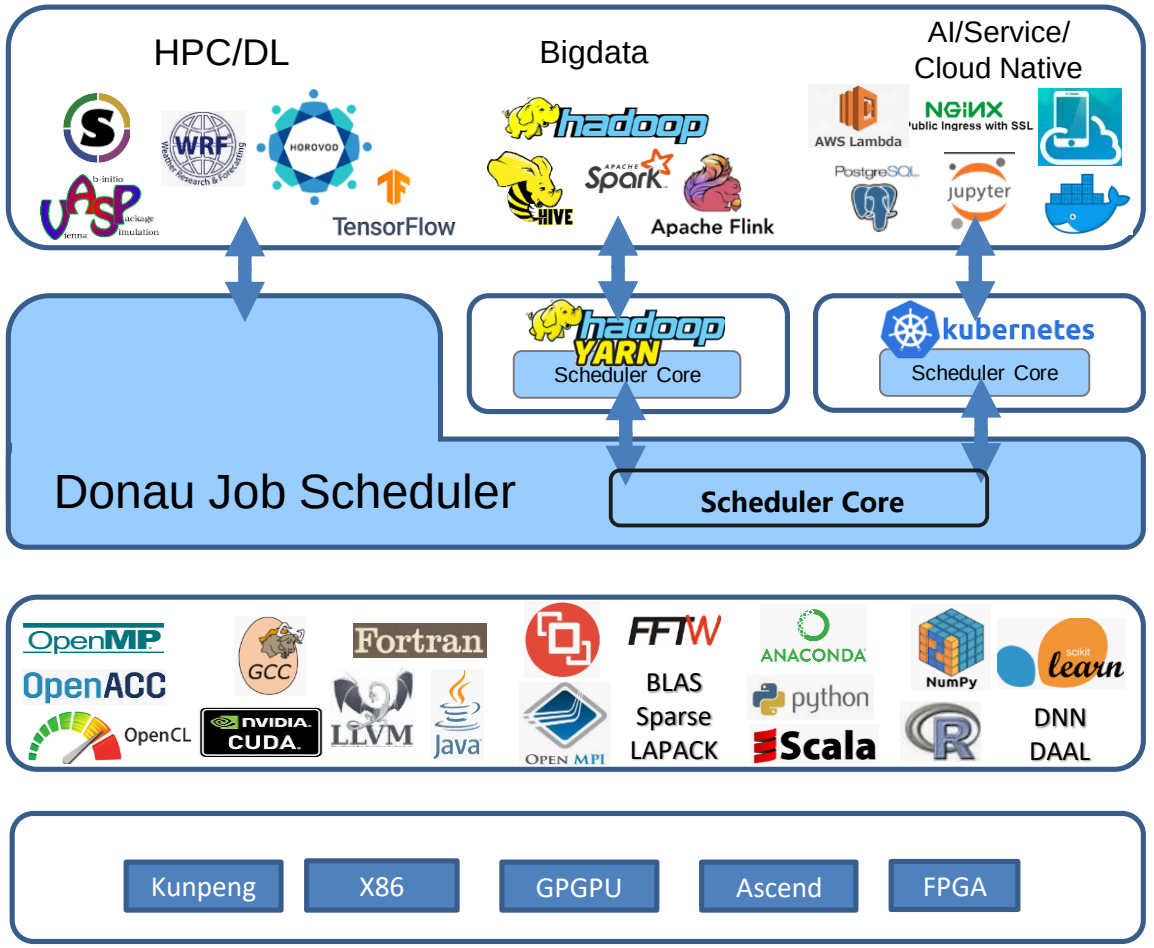
- Donau, a Huawei home-grown job scheduler, launched since 2018
- Target to be a unified scheduler for HPC/Bigdata/AI workload, but no intrusive modification for the application framework
- Extremely high job throughput, target 1M running jobs
- Natively support
  - Accelerator
  - I/O load
  - Container
- Adopt the new technologies in the implementation,
  - Micro-service architecture
  - DRF (Dominate Resource Fairness)
  - MQ & Distributed cache

Donau Scheduler

- Donau – Huawei home-grown HPC scheduler
  - ➢ The scheduler core is an independent module, can replaced the scheduler of Yarn and Kubernetes.

- Scheduling
  - ➢ Natively support elastic allocation
  - ➢ Resource quota based + user priority based fairshare;

- Features:
  - Unified Job/Allocation model: Array Job, MPI job, Bigdata job/Task, TensorFlow job, Service Job, Workflow
  - Natively support container job
  - Accelerator(co-processor) as native resource
  - Resource bursting with IaaS
  - Job-level I/O monitoring & control



HUAWEI

Agenda

HUAWEI

Summary

- Huawei is prompting ARM HPC ecosystem from different dimensions;
- Kunpeng 9x0 ARM-based CPU and Kunpeng mainboard;
- Release HPC Software Suite in the middle of 2020, includes MPI, Compiler, Math Libraries, Scheduler and Management Software;
- Collaborate with Community, Academic and ISV for math libraries, solvers, tool chain and applications.

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

HUAWEI