UCX Development in Huawei

Alex Margolin

UCF Annual Workshop, December 2020

HUAWEI

Outline

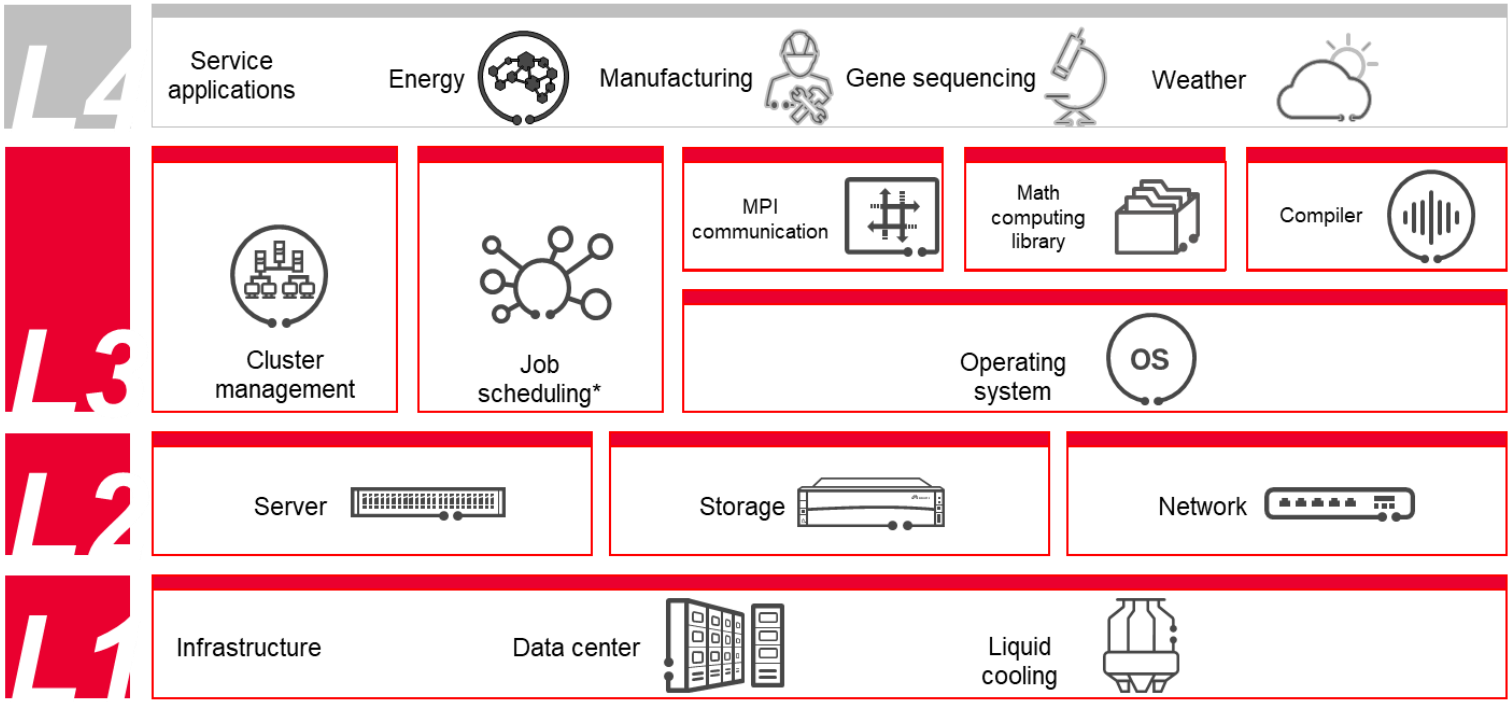
1. Huawei's HPC activities – Cause and effect

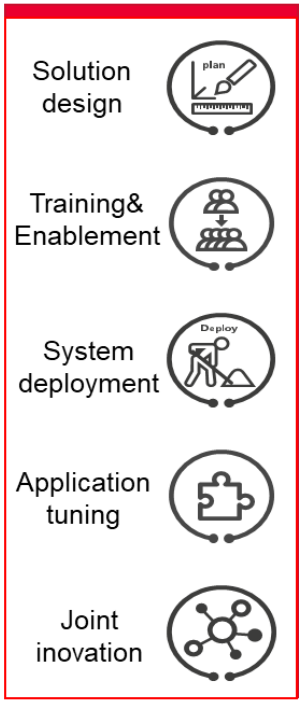
2. Zooming in on UCX (*Today*)

3. Huawei's roadmap for UCX (*Tomorrow*)

4. Teasers from other talks we'll give this week

A Comprehensive HPC Solution – On All Levels





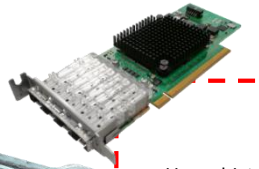
*Supports TaiShan/x86 hybrid scheduling.

A-Z Server Solution by Huawei

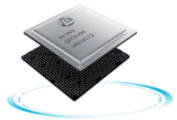
Ascend AI Chip

Huawei Atlas 300 AI Accelerator

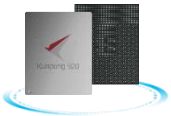
Intelligent NIC Chip



Huawei Intelligent NIC



Kunpeng Processor



Intelligent Management Chip

Intelligent SSD Controller Chip

Huawei ES3000 V5 SSD

HUAWEI | TEL AVIV RESEARCH CENTER

Product Evolution

High computing-efficient

Provide Huawei Kunpeng Processor being compatible with ARM, TaiShan server and efficient computing solutions to help reduce TCO and time to market

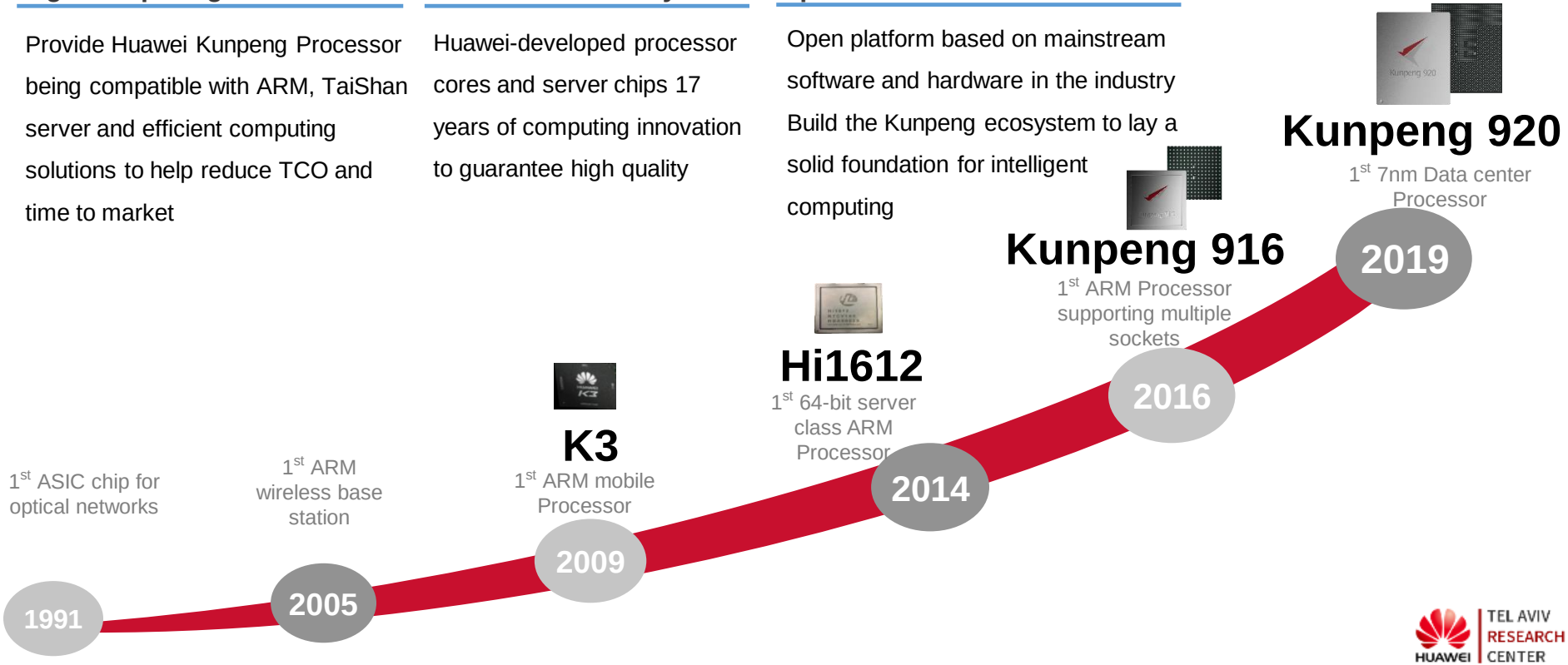
Solid and reliability

Huawei-developed processor cores and server chips 17 years of computing innovation to guarantee high quality

Openness and innovations

Open platform based on mainstream software and hardware in the industry

Build the Kunpeng ecosystem to lay a solid foundation for intelligent computing



Latest CPU – Huawei Kunpeng 920

High Performance

930+ **3x ⬆**

SPECint®_rate_base2006 estimated score

High Bandwidth

Memory bandwidth: **2.4x ⬆**

I/O bandwidth: **1.7x ⬆**

Network bandwidth: **10x ⬆**

High Integration

4 chips in 1

(CPU, south-bridge, NIC, SAS controller)

High Efficient Computing

35% ⬆

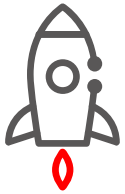
*Tested in Huawei lab, comparison between Kunpeng 920-6426 and last generation Kunpeng 916. Results may vary in different environments

7 nm process | 64 cores | 8 memory controllers | PCIe 4.0 & 100GE

Flexible Form-Factor

2280 Balanced Model	5280 Storage Model	X6000 High-Density Model
		HPC compute node main product
		
For diversified workloads	**5.6 PB per rack**	**10240 cores per rack**
2U Rack Server	4U Rack Server	2U 4-Node Server
2-socket	2-socket	2-socket per node
32*DDR4-2933 MHz	32*DDR4-2933 MHz	16*DDR4-2933 MHz
27*2.5" HDDs or 16*2.5" NVMe SSDs	40*3.5" HDDs	6*2.5" HDDs or NVMe SSDs
CCIX, 8*PCIe 4.0	CCIX, 8*PCIe 4.0	CCIX, 2*PCIe 4.0
GE / 10GE / 25GE	GE / 10GE / 25GE	10GE / 25GE / 100GE RoCE
Air-cooled	Air-cooled	**Air or liquid-cooled**

Breaking Down to Components

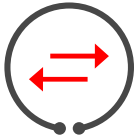


Powerful computing

64-core 2.6 GHz high-performance processor

High memory bandwidth

8 memory channels



High I/O throughput

PCIe 4.0

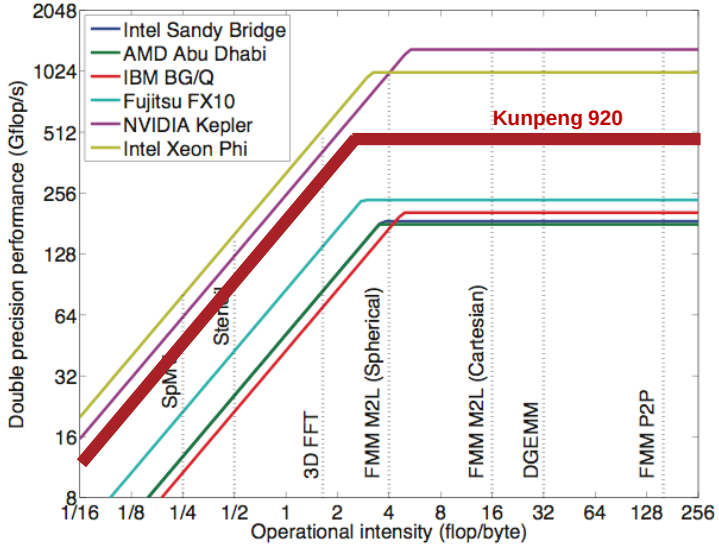
Twice the PCIe 3.0 bandwidth

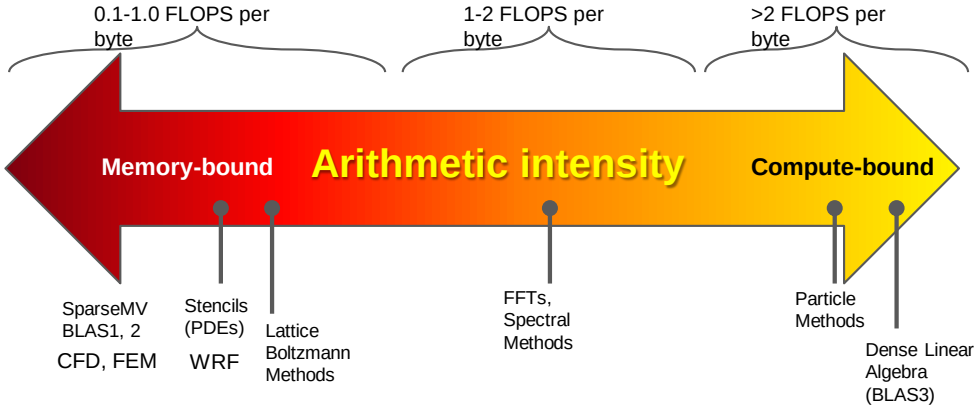
Software and hardware synergized tuning

Huawei-developed compilers and math libraries

Huawei-developed MPI

Classifying by Arithmetic Intensity





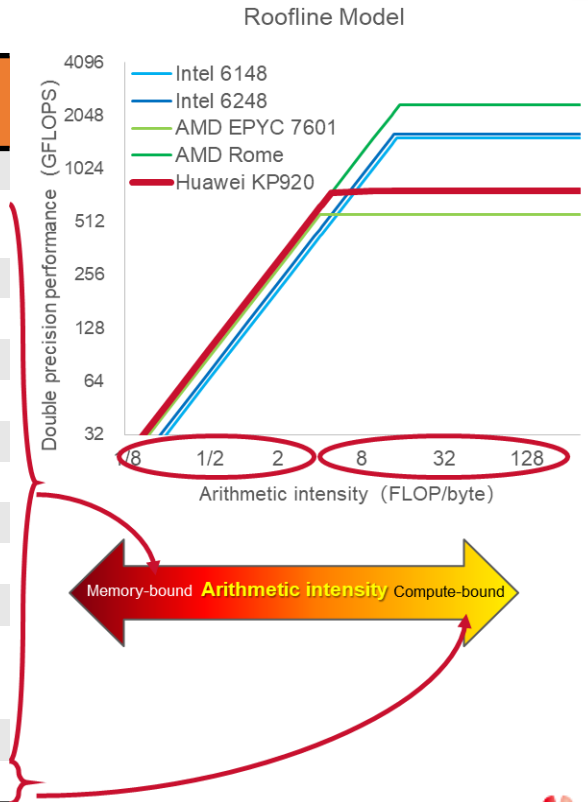
- **Arithmetic intensity** means ratio of (Arithmetic instructions)/(Off-chip memory operands).
- **Lower** arithmetic means **memory-bound**.
- **Higher** arithmetic means **compute-bound.**

Using Roofline model analysis, a large number of HPC algorithms and applications are memory-bound.

TaiShan HPC targets **memory-bound** applications, such as CAE/CFD, weather, life sciences, and oil & gas.

Applications Sample – by Arithmetic Intensity*

Application	Scenario	Numerical Method	Arithmetic Intensity
BCM	CFD	Navier-Stokes	0.14
OpenFoam		Finite Volumes – Finite Element	0.13
Turbine		DNS	0.56
MHD – FDM	Magneto Hydro Dynamics	Finite Difference Method	0.33
MHD - Spectral		Pseudo Spectral Method	0.45
QSFDM	Seismology	Spherical 2.5D FDM	0.46
SEISM3D		Finite Difference Method	0.47
Barotropic	Ocean Circulation Model	Shallow Water Model	0.51
BQCD	High-Energy-Physics	Hybrid Monte-Carlo	0.45 (DP), 0.9 (SP)
B-CALM	Electro-Magnetic Sim.	Finite Difference time-domain	0.3 (SP)
WRF	Weather Forecast model	Stencil code	0.5-1.5
HEPSPEC	SPEC2006 selection for HEP (CERN)	NAMD, DEALII, SOPLEX, POVRAY, OMNETPP, ASTAR, XALANCBMK	>=0.5
Gromacs	Molecular dynamics package	Bennett Acceptance Ratios	<1
KKRNano	Nanotechnology	Korringa-Kohn-Rostoker	4 (DP)



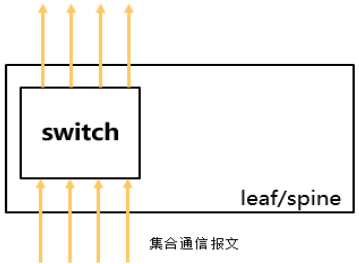
*obviously, this may depend on the input and settings

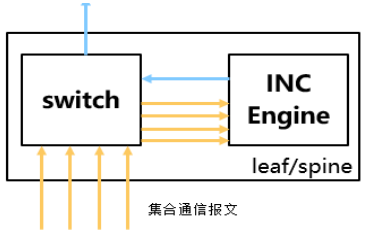
Industry's First RoCE-based Online Computing Solution

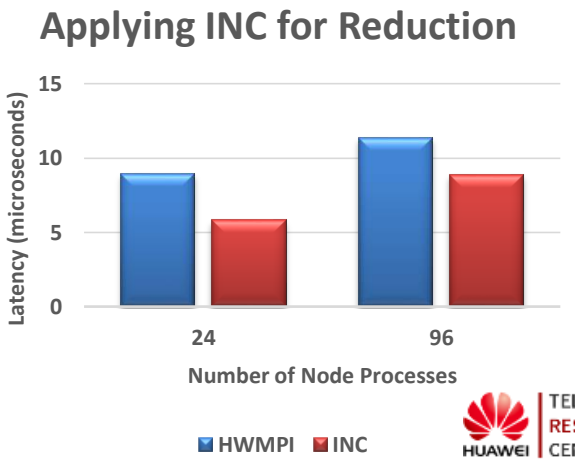
In-network computing (INC):

- Basic principle: add a component in the RoCE switch to offload reduction operations.
- The RoCE-based software and hardware combination solution improves the performance by 30% to 50%.

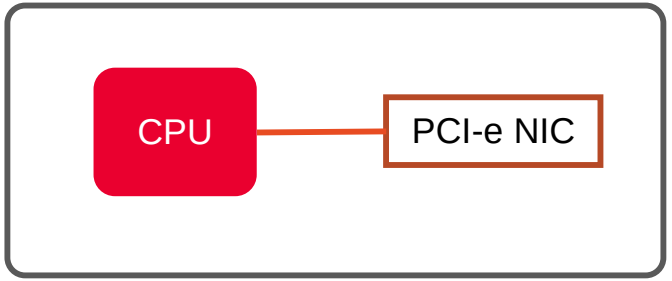
Reduction	PPN	HMPI (RoCE)	HMPI+ INC	INC Improvement
8Bytes	24	8.94	5.81	53.87%
	96	11.36	8.87	28.07%

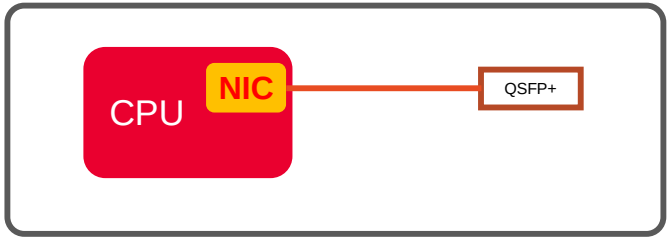






Packets are like sheep… and it's 64 herds!



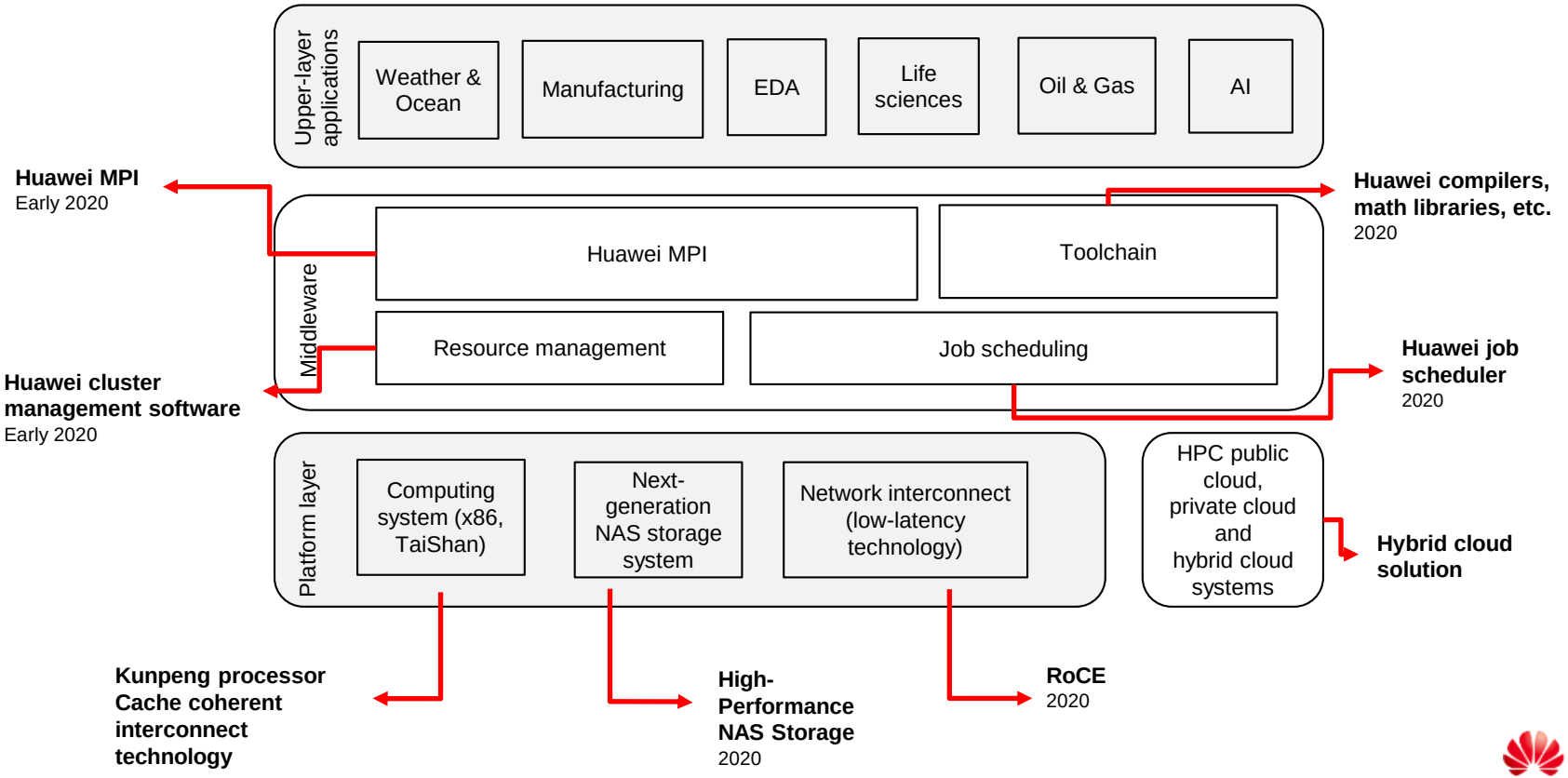


Now #1 in I/O-500

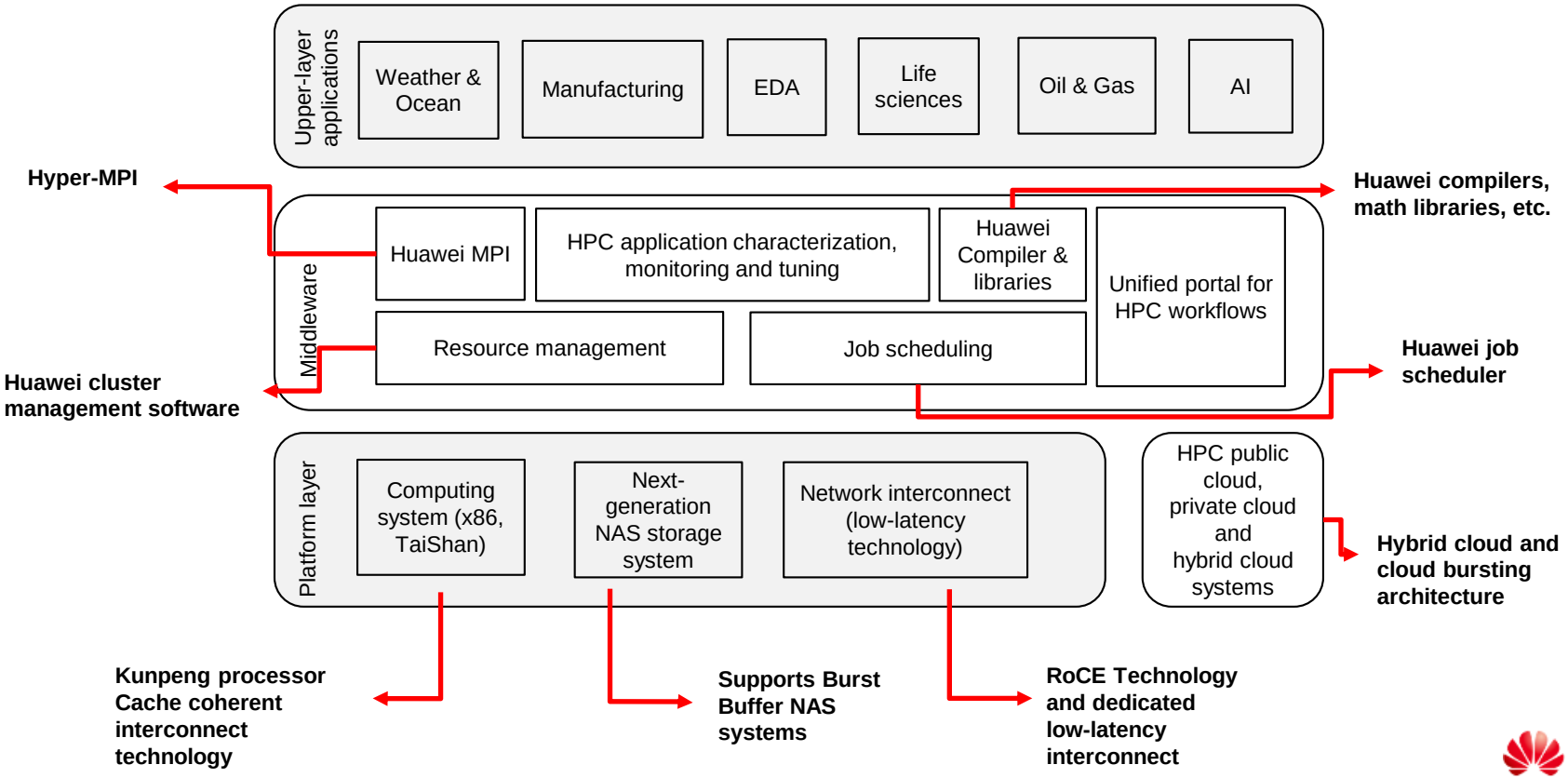
- Capture user-experienced performance
- Reported performance is representative for:
  - applications with well optimized I/O patterns
  - applications with random-like workloads
  - workloads involving metadata small/objects

#	information						io500		
	institution	system	storage vendor	filesystem type	client nodes	client total procs	score	bw	md
								GiB/s	kIOP/s
I/O-500 #1	**Pengcheng Laboratory**	**Pengcheng Cloudbrain-II on Atlas 900**	**Pengcheng Laboratory**	**MadFS**	**255**	**18360**	**7043.99**	**1475.75**	**33622.19**
I/O-500 #2	Intel	Wolf	Intel	DAOS	52	1664	1792.98	371.67	8649.57
I/O-500 #3	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8625	938.95	174.74	5045.33
10-node #1	**Pengcheng Laboratory**	**Pengcheng Cloudbrain-II on Atlas 900**	**Pengcheng Laboratory**	**MadFS**	**10**	**1440**	**1129.75**	**168.43**	**7578.06**
10-node #2	Intel	Wolf	Intel	DAOS	10	420	758.71	164.77	3493.56
10-node #3	TACC	Frontera	Intel	DAOS	10	420	508.88	79.16	3271.49

HPC Solution Overview (when I joined Huawei…)



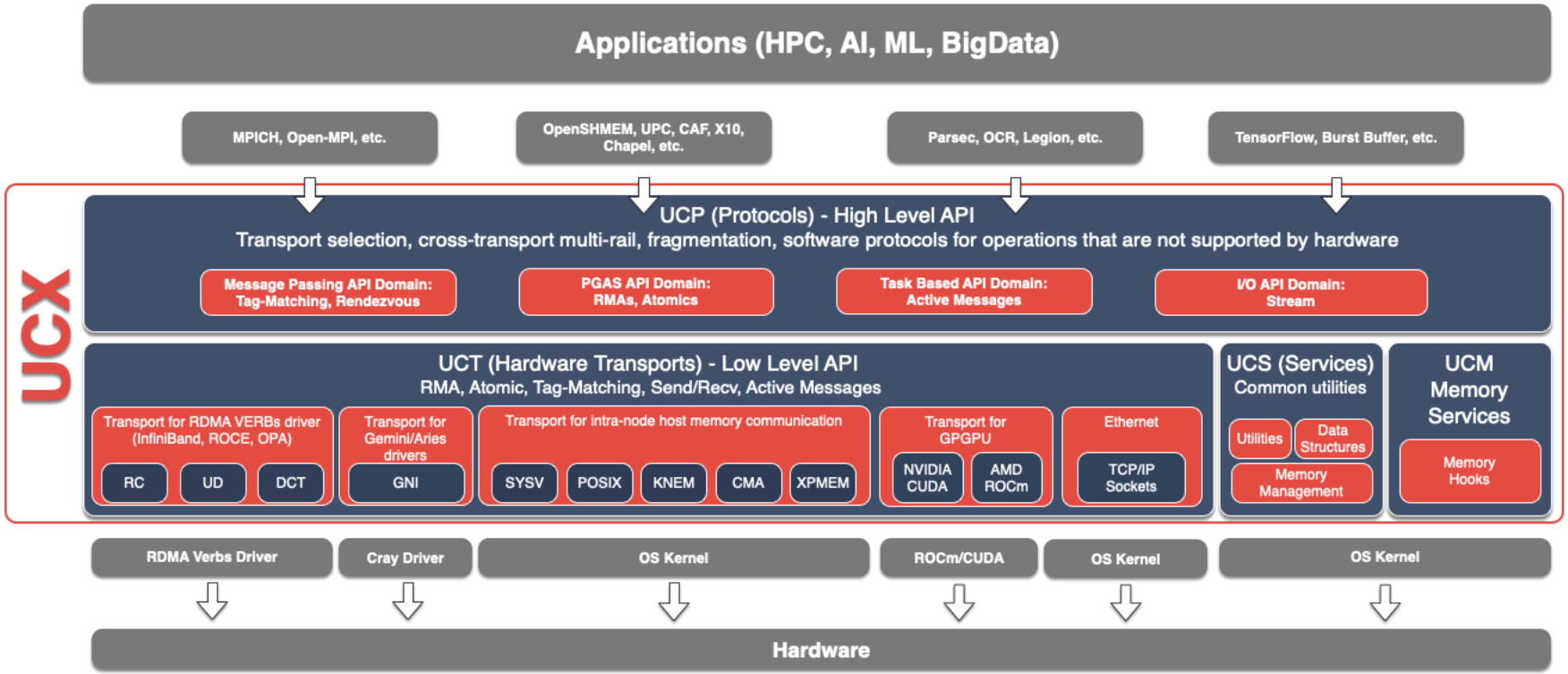
HPC Solution Overview Today



Outline

1. Huawei's HPC activities – Cause and effect
2. Zooming in on UCX (*Today*)
3. Huawei's roadmap for UCX (*Tomorrow*)
4. Teasers from other talks we'll give this week

Gentle Reminder



UCS (Services)

Pointer-Array

- Added *locked-pointer-array*, for thread-safety where applicable
- **~30-35%** increase in iteration speed of *foreach()*: **1.46 ns → 1.02 ns**
- **10%** faster iteration on x86 (test-specific…), thanks to prefetching trickery

Usage? In UCG*: each column represents a context, and has a pointer-array to hold messages (equiv. to Tag-matching).

Slot 0	Slot 1	Slot 2	Slot 3
A	C	D	
B		E	
		F	

Statistics

- Apply filters to reported stats (*over UDP)
- To be continued…

Misc.: timer queue, aligned realloc., GCC fixes, etc.

UCT (Transports)

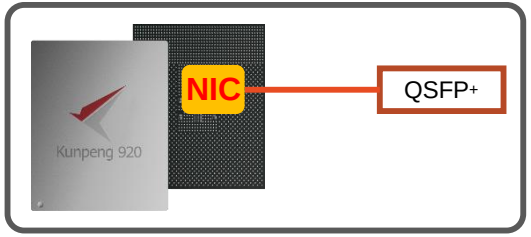
SPOILER ALERT

One-to-many communication

- Expand UCX's P2P focus to cover other types of communication
- Lots of changes are required for this: AM ID range, "exposing" internals, etc.

Kunpeng's RoCE support

- Basically calls rdma-core, but there are some specifics (not public yet)



RoCE Reachability issues

Suppose you have 2 RoCE ports: do you…

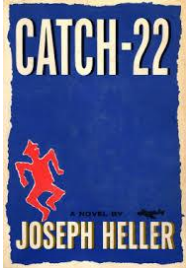
a. set both IP addresses on the same subnet? (how to choose TX port? + Socket-Direct is unhappy)

b. set IP addresses on separate subnets? (some ports can't talk to other ports!)

Workaround: ask for RoCEv1 (no direct way, use UCX_IB_GID_INDEX=0)

Not a workaround: Link aggregation (LAG)

Solution: have UCX choose RoCEv1 (PR #5581)



UCP (Protocol)

One-to-many communication

- Lots of changes are required for this: transport selection, extra API parameters, etc.
- **Not much else, since we mostly use UCT directly – see next slide…**

Q: Why use UCT directly?!

A: For several reasons, incl. mostly overhead considerations and too little control when using UCP. More on this during the UCG talk, later this week.

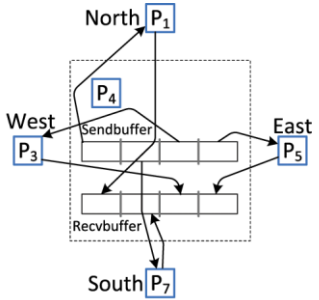
UCG (Groups)

- Huawei's choice for collective operations.
- Will (eventually) support any collective type.

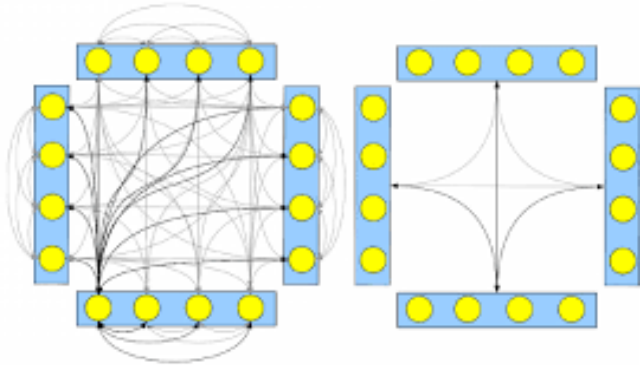
  *yes, including **`MPIX_Neighbor_alltoallw_init()`**
- Shamelessly (ab)uses UCX internals to work faster (incl. internal UCT stuff).
- Lives at https://github.com/openucx/xucg (X for eXperimental, a stable version exists)

SPOILER ALERT

- To be continued…



(a) Traditional neighborhood alltoall



Open-MPI support

- Is there anything missing? **Yes, we think so.**

Code consolidation among UCX components (for starters)

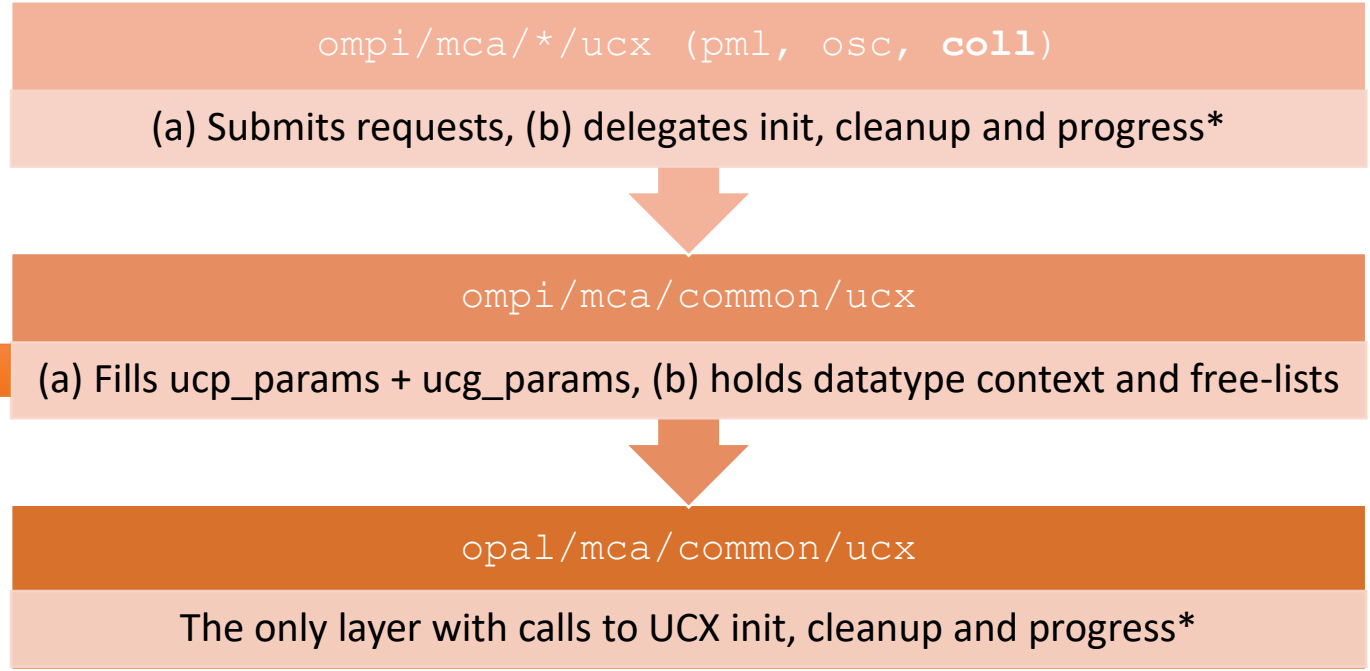
- ./ompi/mca/coll/ucx
- ./ompi/mca/osc/ucx
- ./ompi/mca/pml/ucx
- ./ompi/mca/common/ucx
- ./opal/mca/common/ucx
- ./oshmem/…
- *At least 2-3 additional components are in the planning!*

- Better integration with the rest: MPI_T, hints when used by other components…

Open-MPI Component Consolidation

Why? Because on x6000 – we have 256 cores (x2 workers, x2 QPs - no good)

How?



Another feature: the same non-contiguous datatype does not get duplicated across PML and COLL MCAs.

*except during blocking calls on OMPI/mca level

Open-source policy

- Most of what we do is contributed… but not everything gets upstream.
- * Not just UCX / MPI – also KNEM, Spack, Open-PMIx / prrte, etc.

- There are a few exceptions:

1. Some hardware-specific code, especially for HW not on the market yet.

2. Integration code with some of Huawei's proprietary software (e.g. Storage)

3. Code that didn't pass Huawei's code quality checklist **YET**…. takes time!

Outline

1. Huawei's HPC activities – Cause and effect
2. Zooming in on UCX (*Today*)
3. Huawei's roadmap for UCX (*Tomorrow*)
4. Teasers from other talks we'll give this week

Integration with other parts of Huawei's HPC solution

- When you have a hammer…

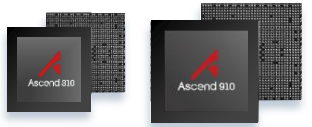
Taishan Server

CloudEngine (Switch)

1. Working to get UCX adopted in other products (P2P and collectives).

OceanStor (Storage)

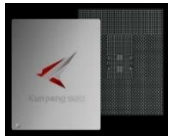
2. Working with the compiler team and the performance teams to optimize the build on our processing elements (not just CPUs).



Ascend 310 & 910 (AI)

SPOILER ALERT

3. Working to accelerate UCX features we need



Kunpeng 920 (CPU)

IN200 (100GbE NIC)

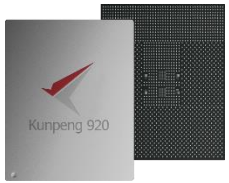
- This goes both ways – <u>UCX could use more input/hints</u> !

Example: Kunpeng-specific optimizations

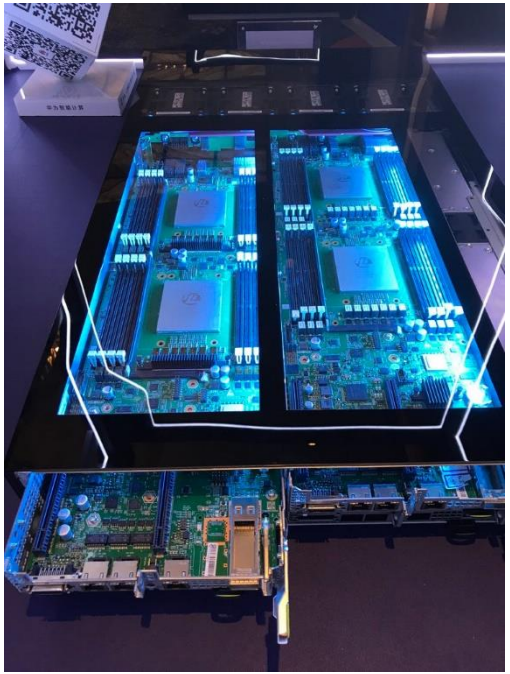
The challenge: up to 64 cores per CPU, up to **256 per host**!

This effects:

- algorithm selection,
- resource constraints, (*remember OMPI code consolidation?)
- transport selection,
- …
- **Application performance.**



X 4 =



Top features for high PPNs:

1. Shared-memory comm. enhancement (P2P and collectives),
2. Finer-grained topology awareness (even within a CPU),
3. Memory footprint reduction,
4. Good old testability at scale.

Concrete Example: Kunpeng-specific optimizations

Proposal: (external) user-space locking library

- Requires some research, to choose the best library to use, for example (publications):
- "*Compact NUMA-aware Locks*" by Dice and Kogan (EuroSys '19)
- "*Scalable and practical locking with shuffling*" by Kashyap et al. (SOSP '19)
- … depends a lot on the architecture:

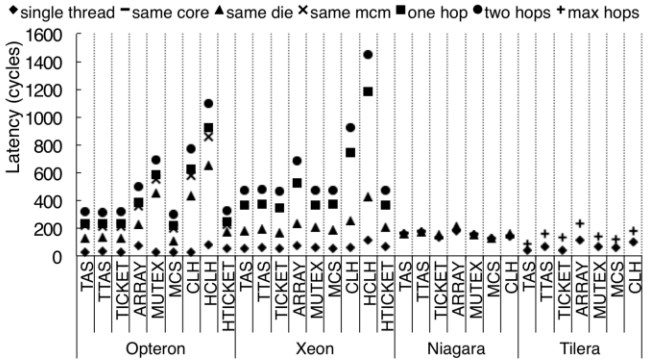


Figure 6: Uncontested lock acquisition latency based on the location of the previous owner of the lock.

System	Opteron (2.1 GHz)				Xeon (2.13 GHz)			Niagara (1.2 GHz)		Tilera (1.2 GHz)	
Hops	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: Compare & Swap (C), Fetch & Increment (F), Test & Set (T), Swap (S)											
Operation	all	all	all	all	all	all	all	C F T S	C F T S	C F T S	C F T S
Modified	110	197	216	296	120	324	430	71 108 64 95	66 99 55 90	77 51 70 63	98 71 89 84
Shared	272	283	312	332	113	312	423	76 99 67 93	66 99 55 90	124 82 121 95	142 102 141 115

Optimizations Galore!

- New ways to transfer data (even with existing interconnect HW),
- New ways to overlap computation and communication,
- New ways to detect and optimize common patterns,
- New ways to help applications.

A Bit about my team in Israel…

Appears to be a software engineering team, focused on networking…

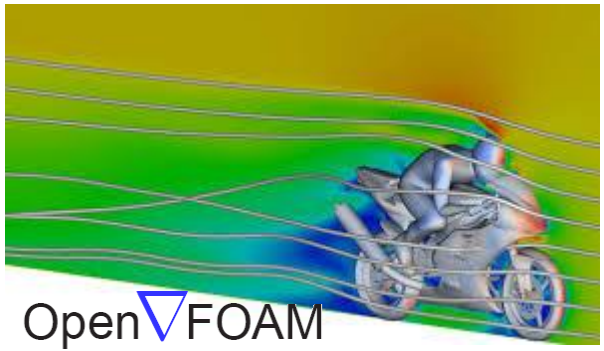
Our secret weapon?

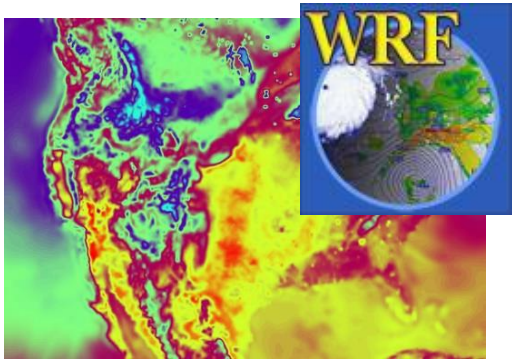
A Bit about my team in Israel…

Appears to be a software engineering team, focused on networking…

Our secret weapon?

- Half of my team are scientists (not computer scientists…), incl. professors, experts on computational-*: CFD, atmospheric models, molecular dynamics, etc.





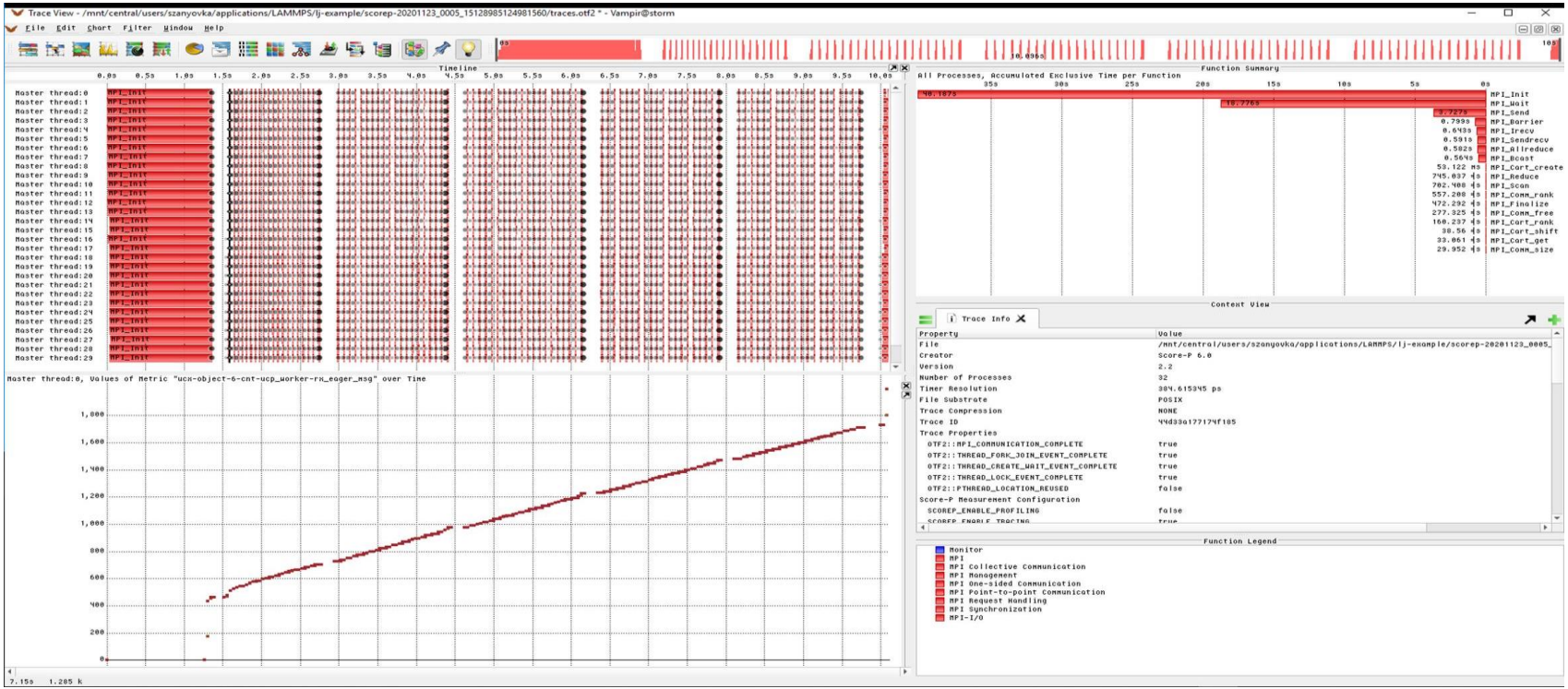
P.S. SC20 Everywhere we are | more than hpc. is preaching to the choir… look for us at SCI-TECH FORUM and 2020 OpenFOAM® Journal

HUAWEI | TEL AVIV RESEARCH CENTER

Outline

1. Huawei's HPC activities – Cause and effect
2. Zooming in on UCX (*Today*)
3. Huawei's roadmap for UCX (*Tomorrow*)
4. Teasers from other talks we'll give this week

UCX counters in Score-P and Vampir (Day 3, 10:40 CT)



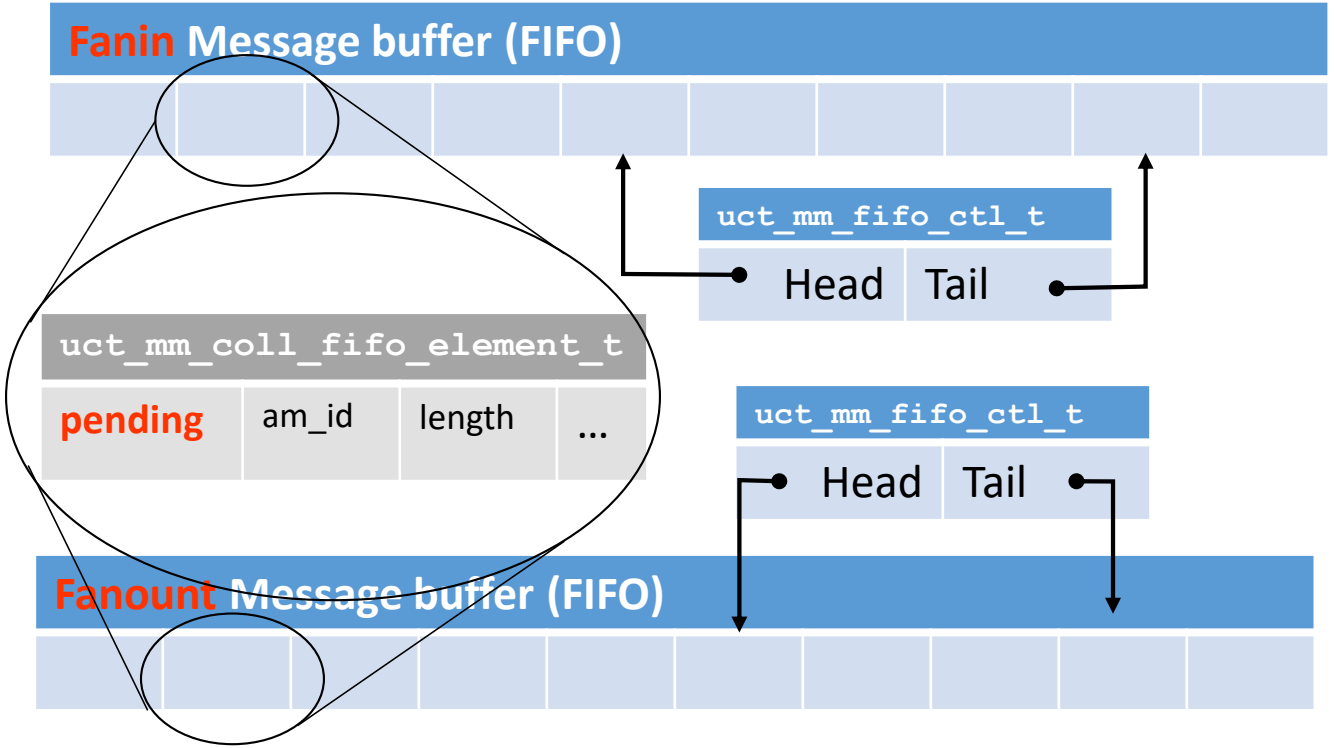
One-to-many UCT transports (Day 4, 9:00 CT)

4 queues needed:

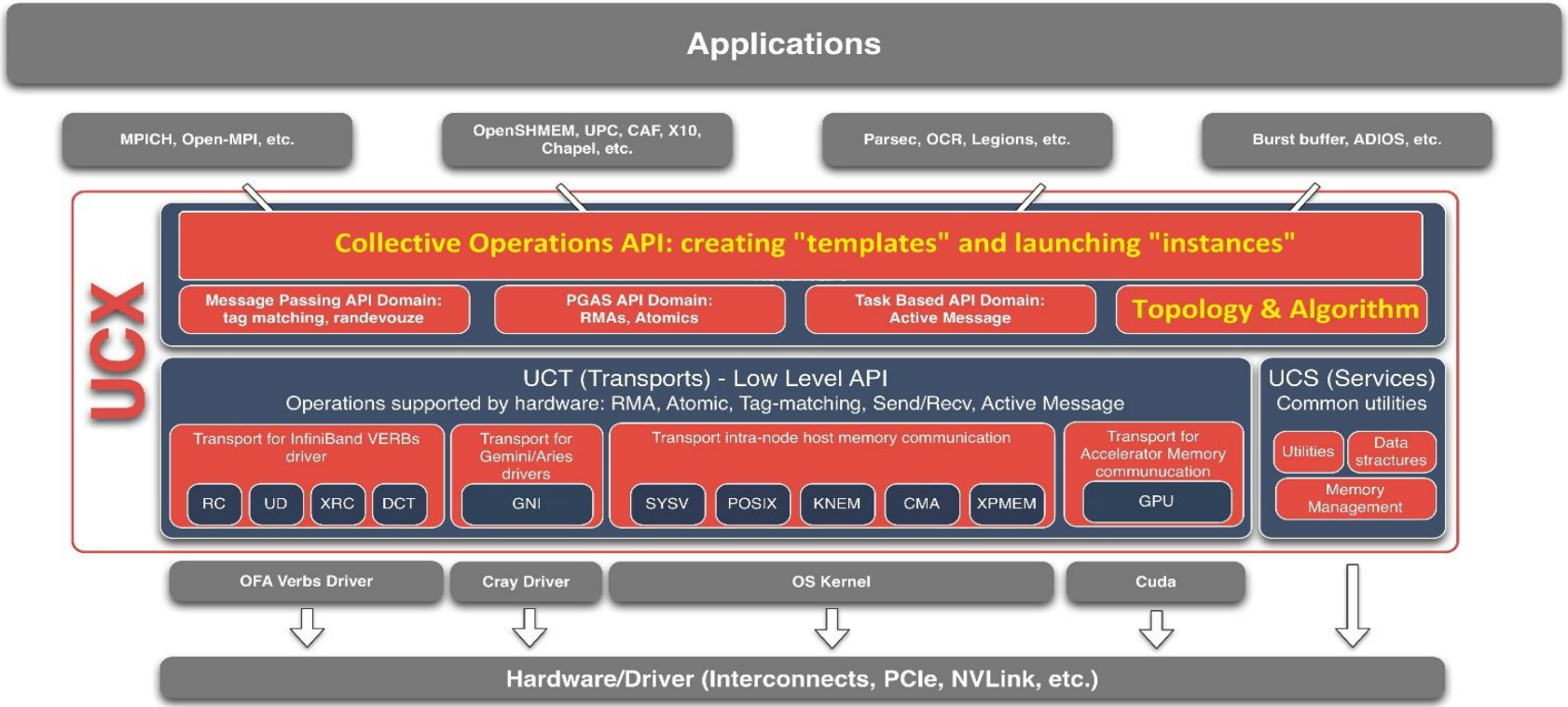
1+2. The existing P2P queues, for control messages (e.g. Rendezvous).

3. Fanin, for collectives like reduce or gather.

4. Fanout, for collectives like bcast and scatter.



Until UCC is available - UCG status update (Day 4, 10:00 CT)



www.huawei.com