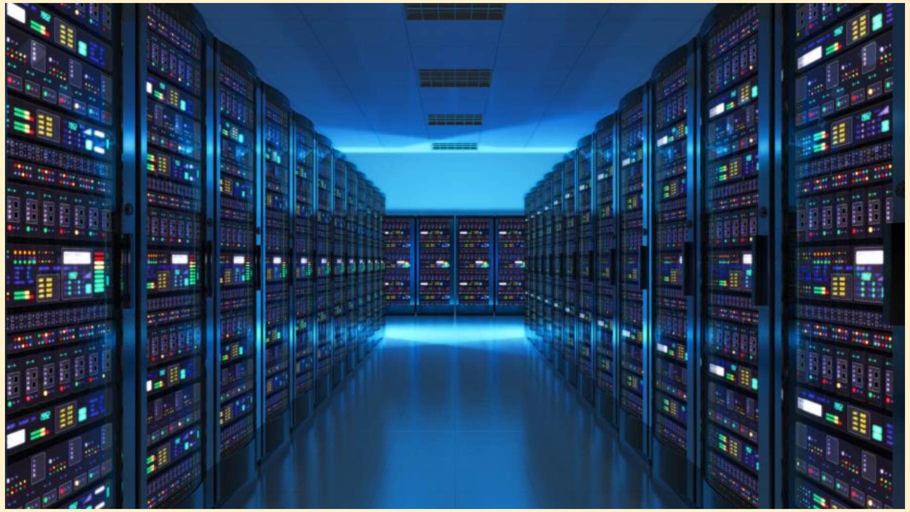
The Seventh ACM/IEEE Symposium on Edge Computing
Seattle, WA, December 5-8, 2022

Position Paper: Renovating Edge Servers with ARM SoCs

Mengwei Xu, Li Zhang, Shangguang Wang
Beijing University of Posts and Telecommunications

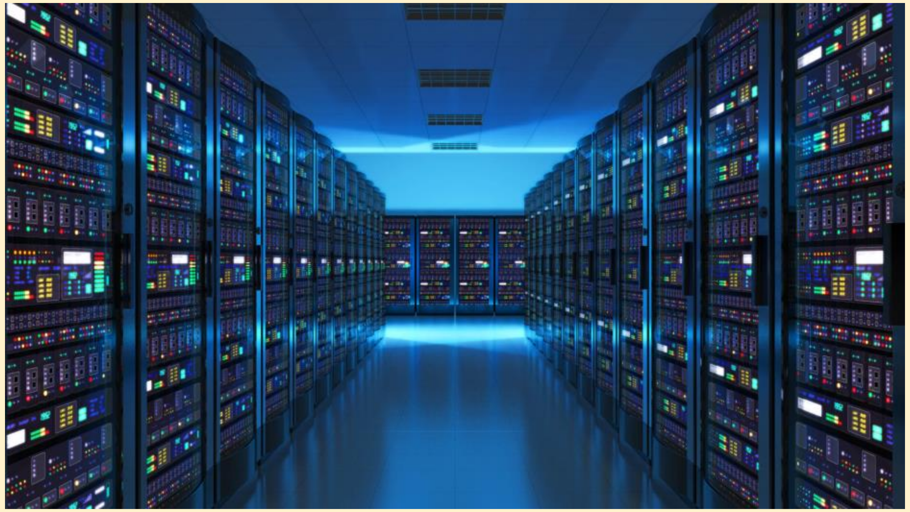
Cloud vs Edge



Cherry-picked, often rural areas

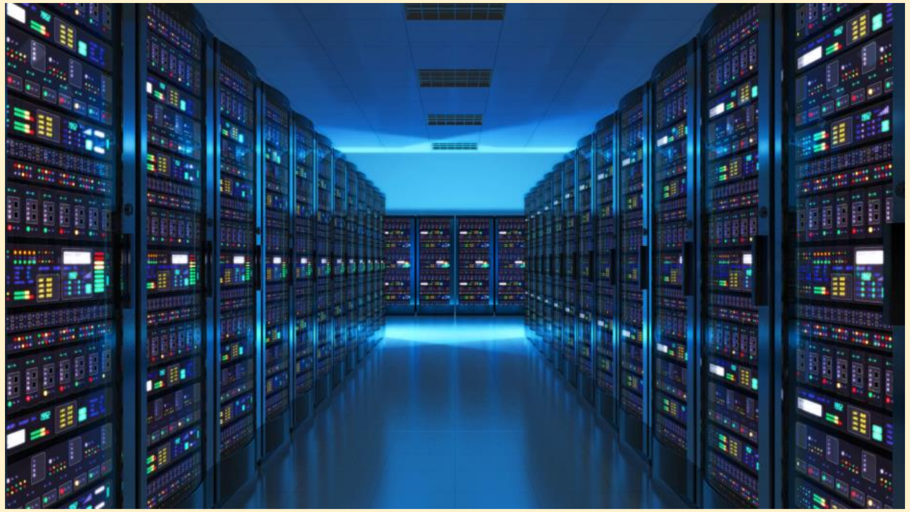
Near-population, urban areas

Cloud vs Edge



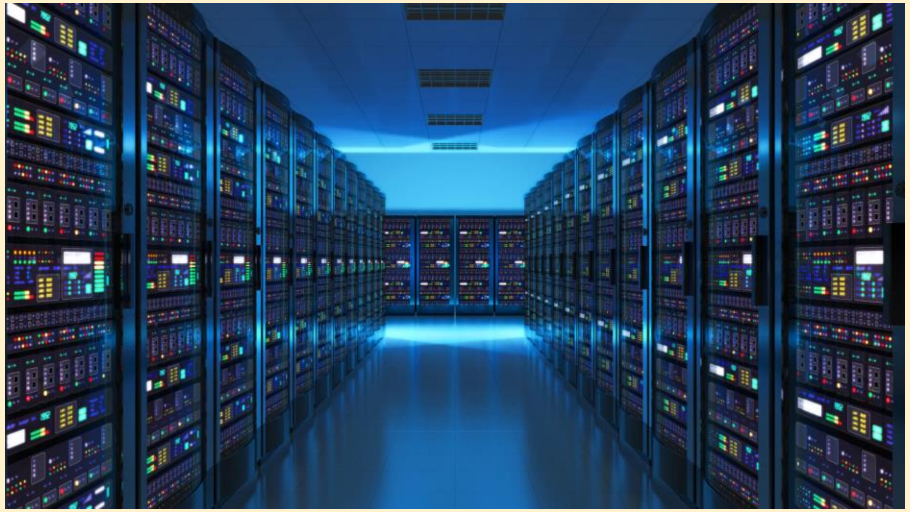
Large, scalable **<u>Space</u>** **Limited, unscalable**

Cloud vs Edge



Large, scalable	**Space**	Limited, unscalable
Abundant, cheap	**Power Supply**	Constrained, expensive

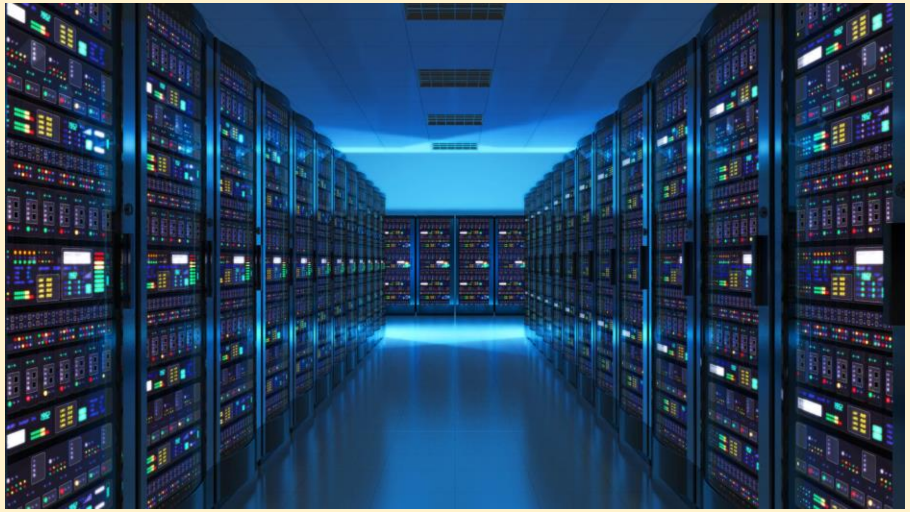
Cloud vs Edge



Cloud		Edge
Large, scalable	**Space**	**Limited, unscalable**
Abundant, cheap	**Power Supply**	**Constrained, expensive**
Powerful, mature	**Cooling Facility**	**Wimpy or even doesn't exist**

Pei, Qiangyu, et al. "CoolEdge: hotspot-relievable warm water cooling for energy-efficient edge datacenters." *ASPLOS* 2022.

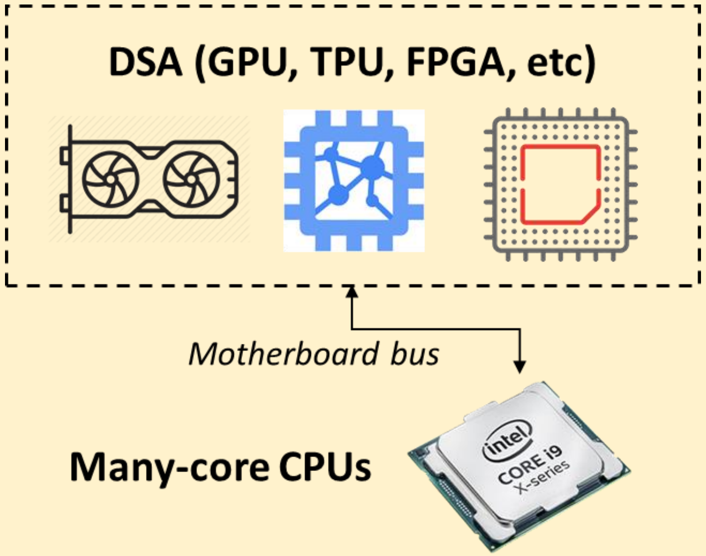
Cloud vs Edge



Xu, Mengwei, et al. "From cloud to edge: a first look at public edge platforms." *IMC* 2021.

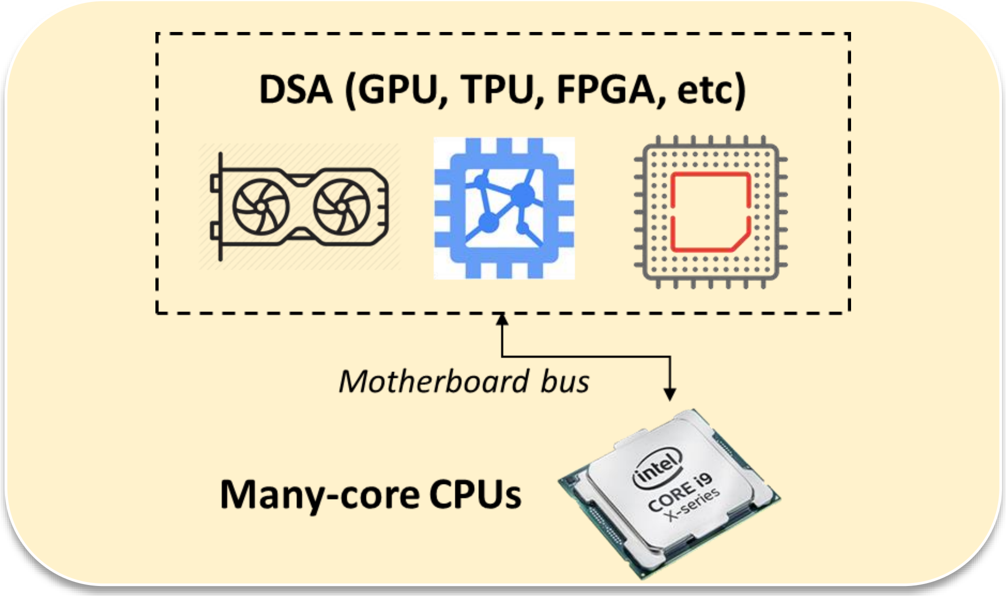
Large, scalable	**Space**	**Limited, unscalable**
Abundant, cheap	**Power Supply**	**Constrained, expensive**
Powerful, mature	**Cooling Facility**	**Wimpy or even doesn't exist**
Various types, stable	**Workloads**	**Domain-specific, highly variational**

Cloud vs Edge



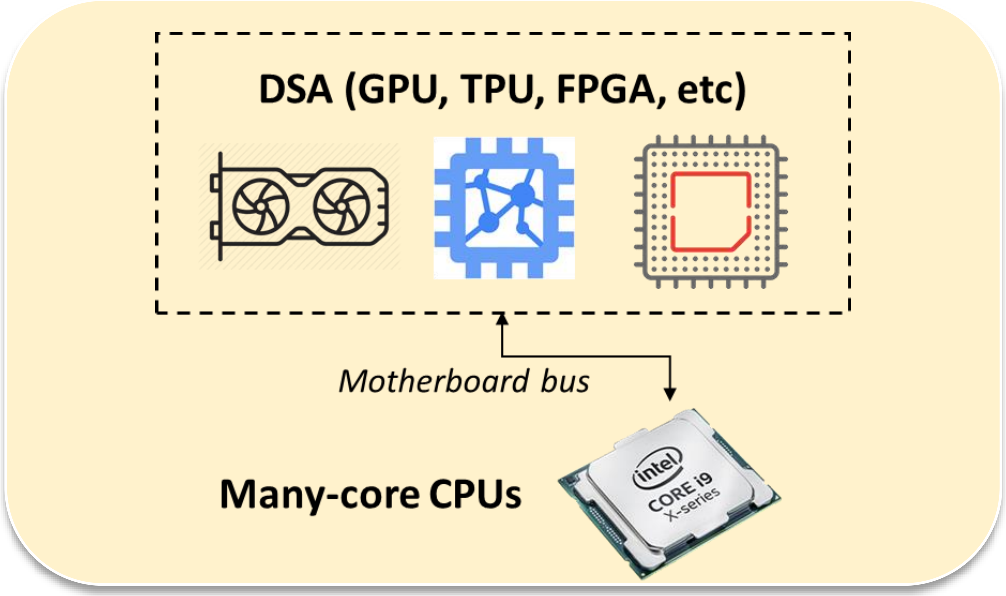
Large, scalable	Space	Limited, unscalable
Abundant, cheap	Power Supply	Constrained, expensive
Powerful, mature	Cooling Facility	Wimpy or even doesn't exist
Various types, stable	Workloads	Domain-specific, highly variational

Cloud vs Edge



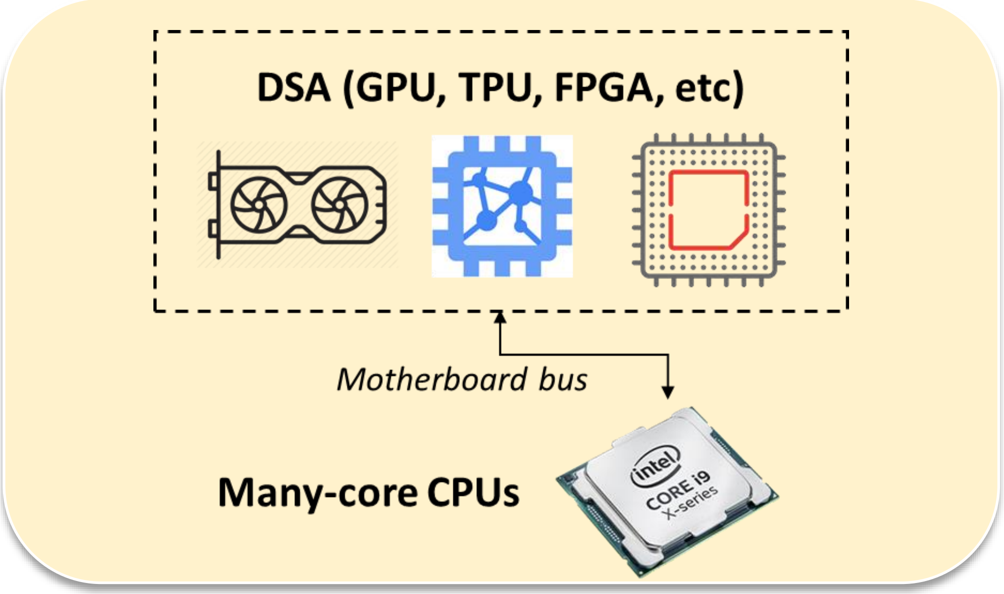
Large, scalable	**Space**	**Limited, unscalable**
Abundant, cheap	**Power Supply**	**Constrained, expensive**
Powerful, mature	**Cooling Facility**	**Wimpy or even doesn’t exist**
Various types, stable	**Workloads**	**Domain-specific, highly variational**

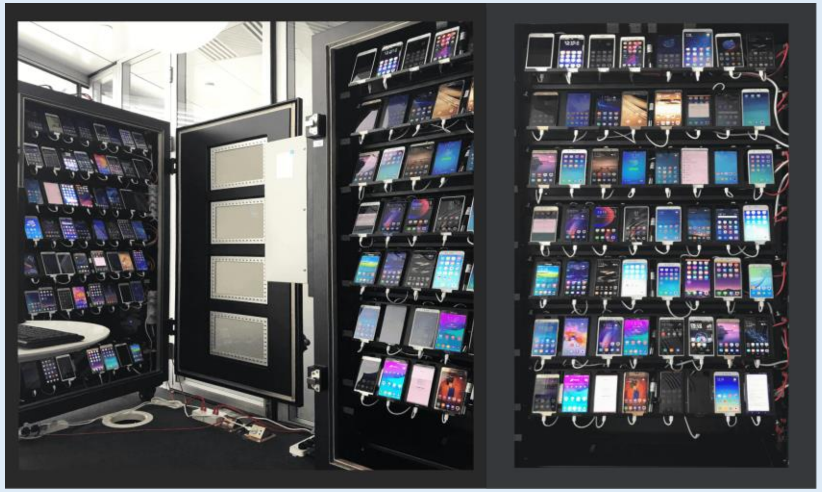
Cloud vs Edge



A smartphone!

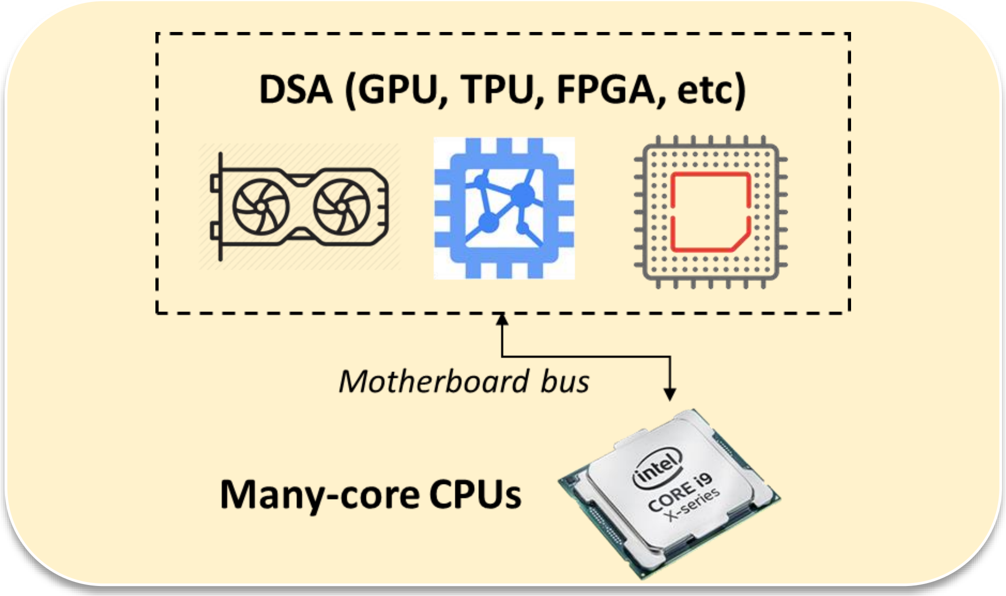
Cloud vs Edge

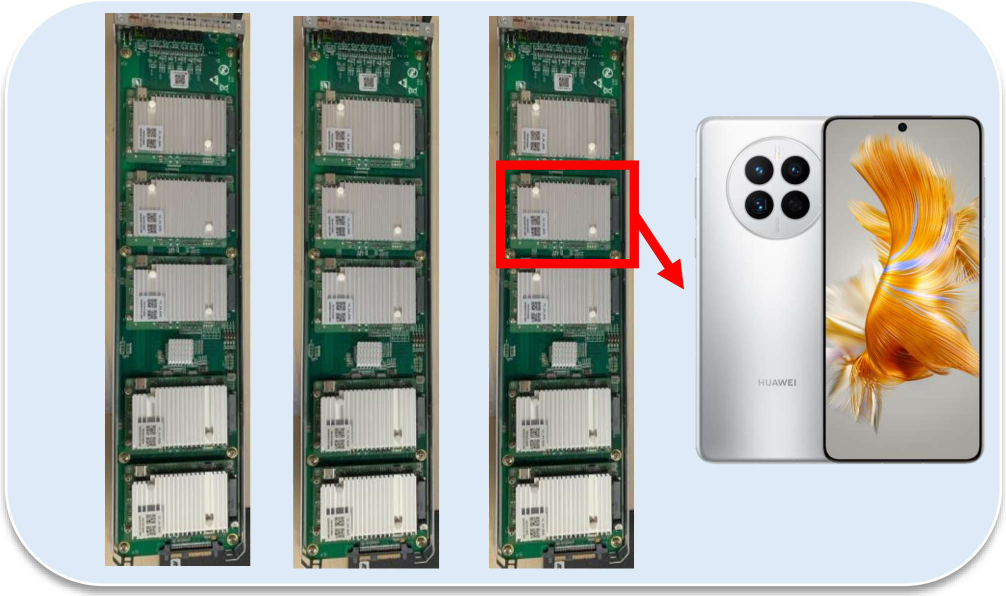




Many smartphones!

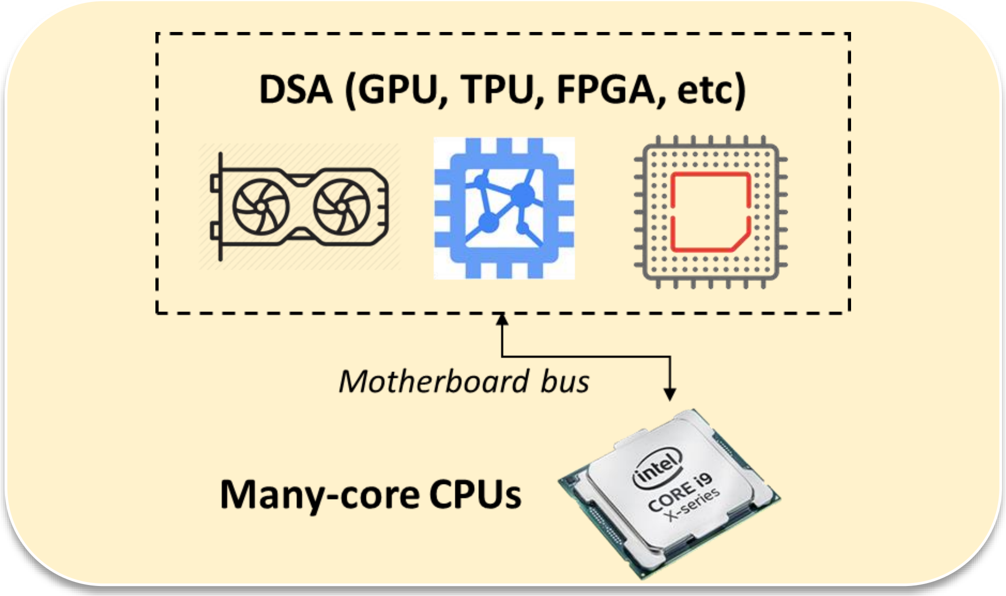
Cloud vs Edge

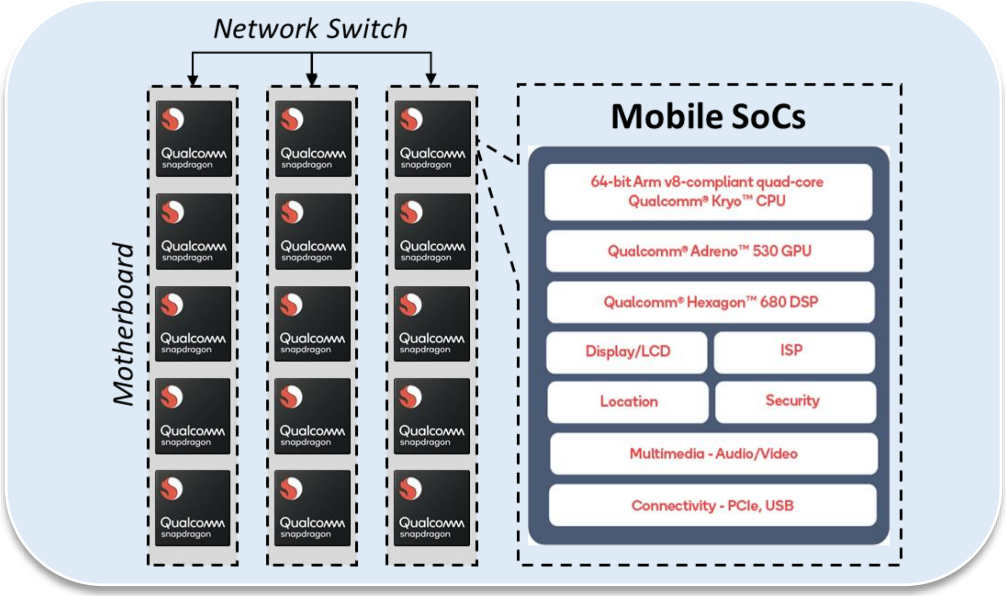




Many SoCs!

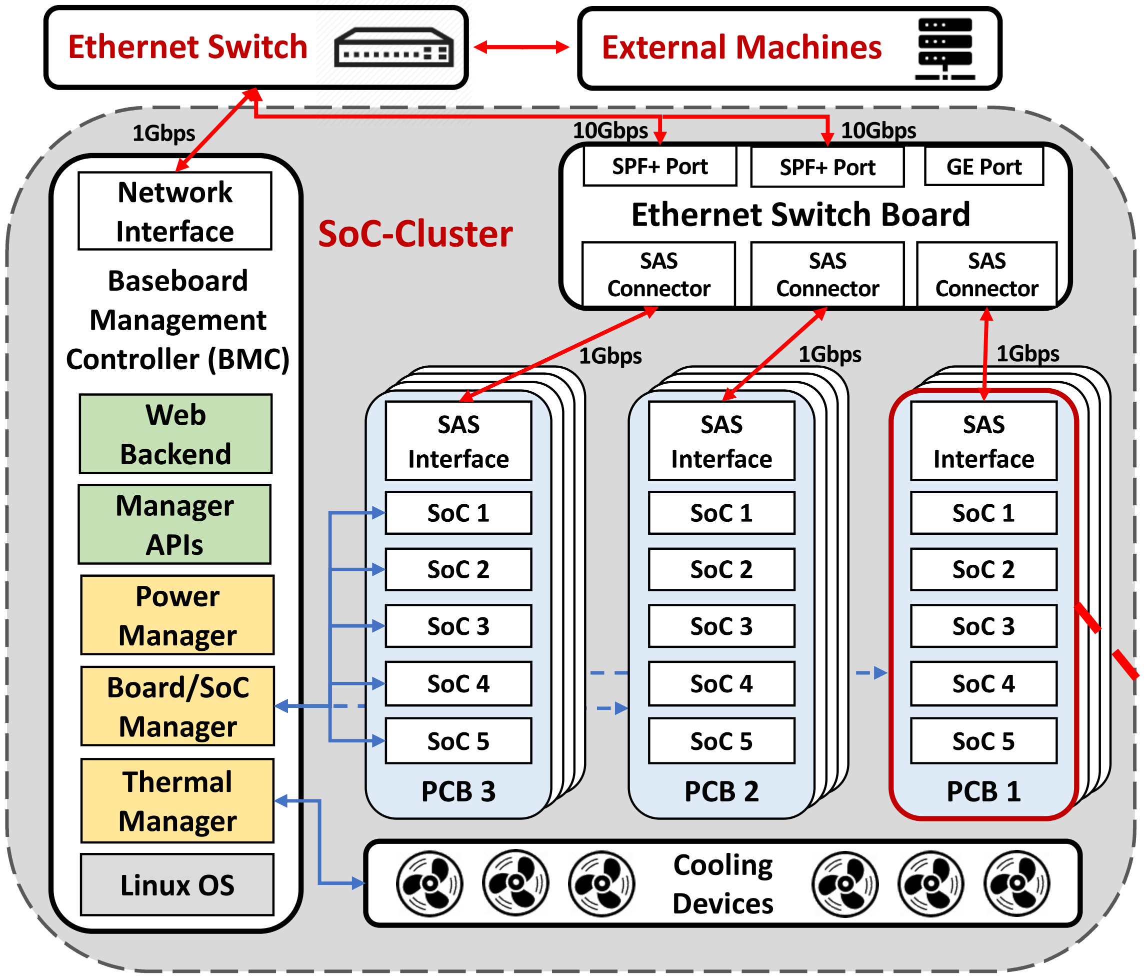
Cloud vs Edge



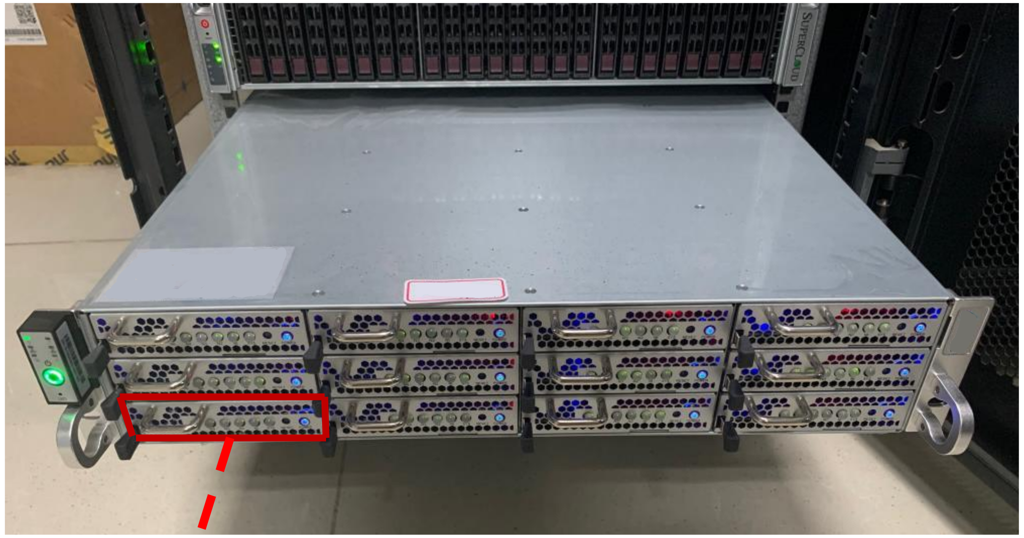


Monolithic vs. Decentralized

Our proposal of SoC-Cluster



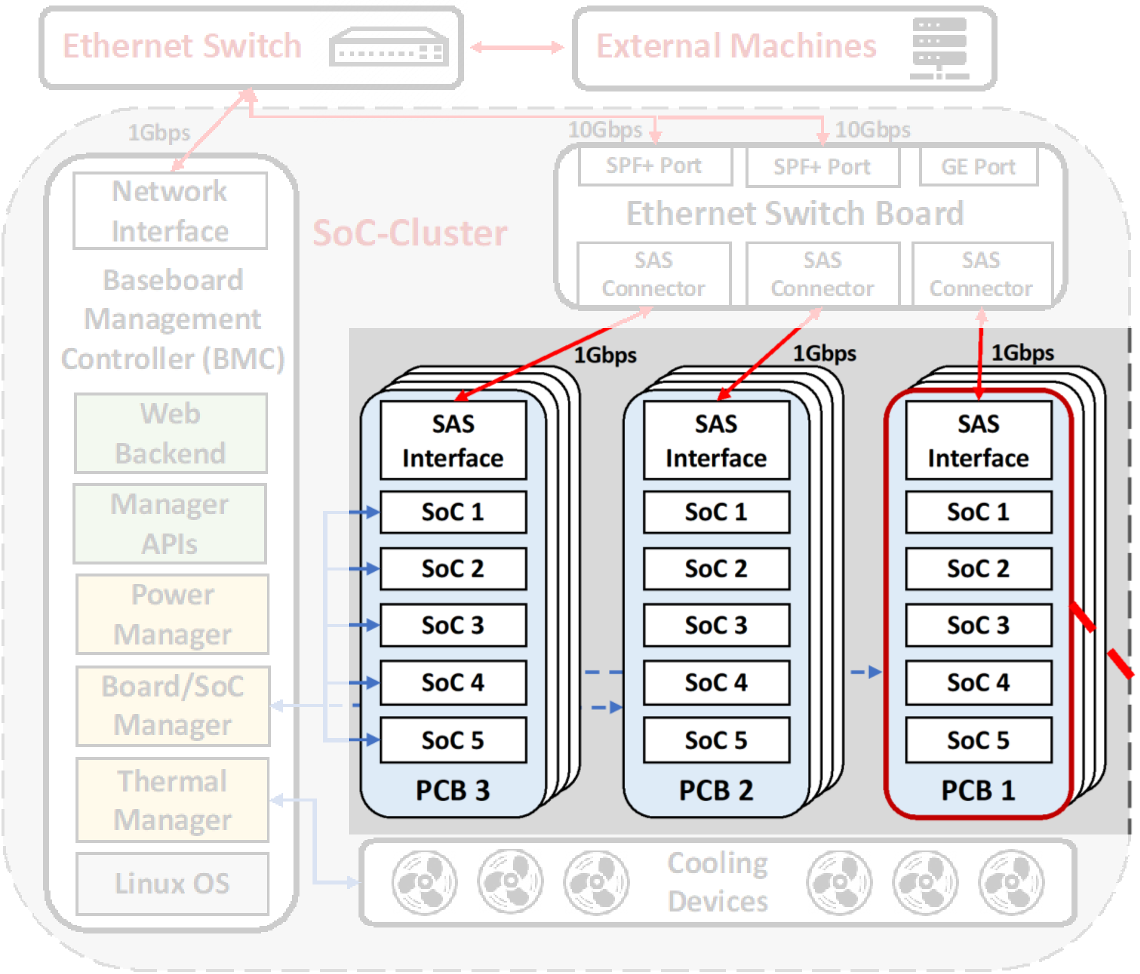
(a) The conceptual architecture.



(b) The overview of a physical server.

(c) The internal PCB board with 5 SoCs.

Our proposal of SoC-Cluster

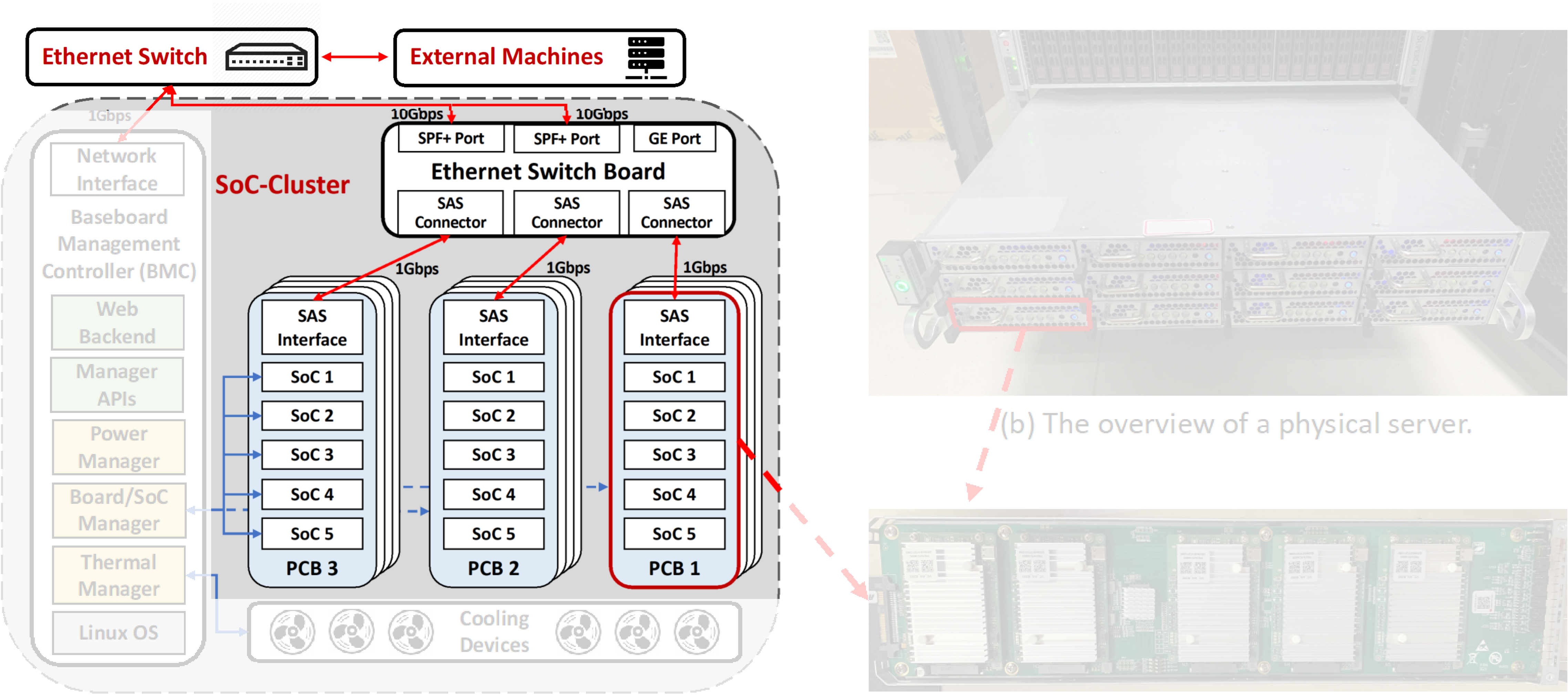


(a) The conceptual architecture.

(b) The overview of a physical server.

(c) The internal PCB board with 5 SoCs.

Our proposal of SoC-Cluster

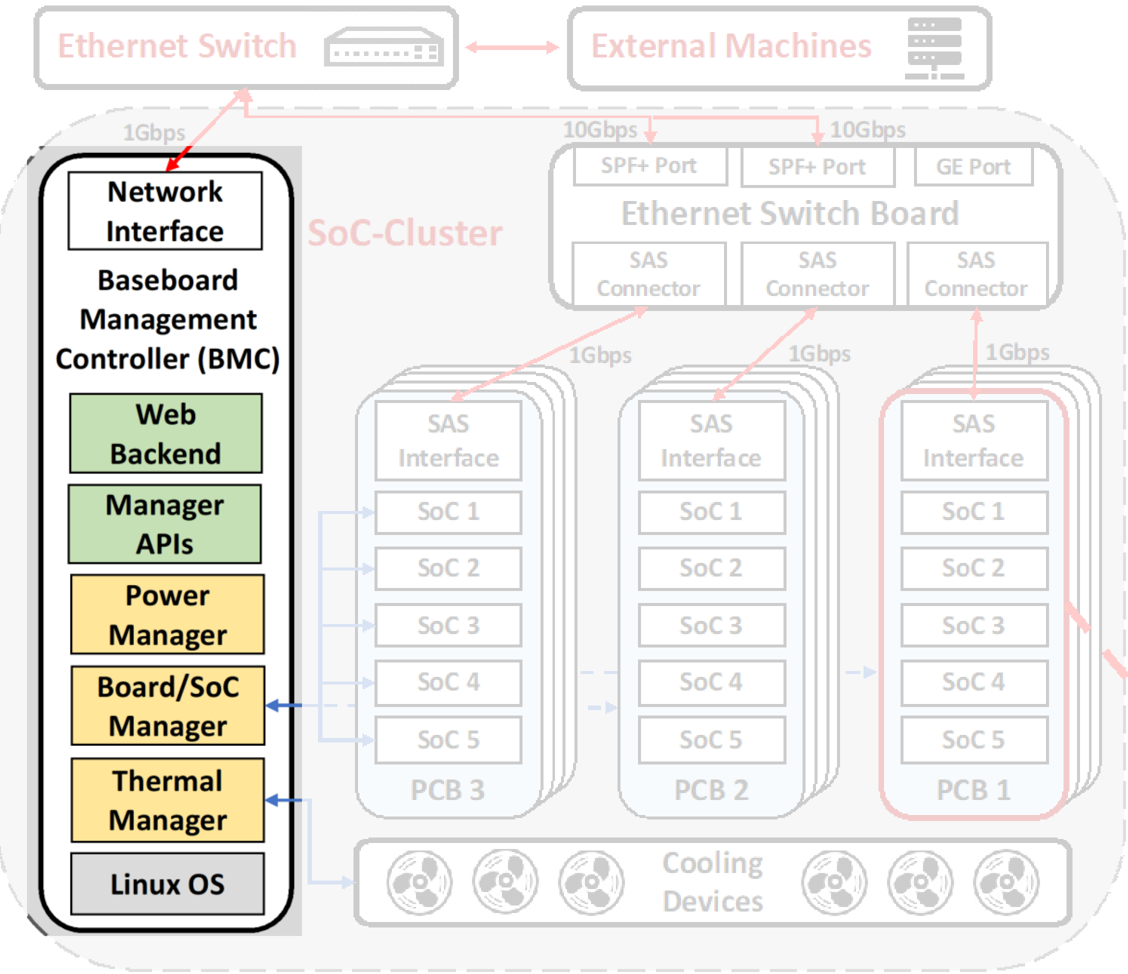


(a) The conceptual architecture.

(b) The overview of a physical server.

(c) The internal PCB board with 5 SoCs.

Our proposal of SoC-Cluster



(a) The conceptual architecture.

(b) The overview of a physical server.

(c) The internal PCB board with 5 SoCs.

北京邮电大学
Beijing University of Posts and Telecommunications

Our proposal of SoC-Cluster

Server in 2U rack

	SoC-Cluster (60x Snapdragon 865)	**Conventional GPU Server (4 x NVIDIA V100)**
CPU	400 cores	< 100 cores
Accelerators	50x Adreno GPUs ➢ 50 TFLOPS 50x Hexagon DSPs ➢ 750 TOPS	400 TFLOPS
Memory	600GB	< 200GB
Disk (Flash)	10TB	10TB

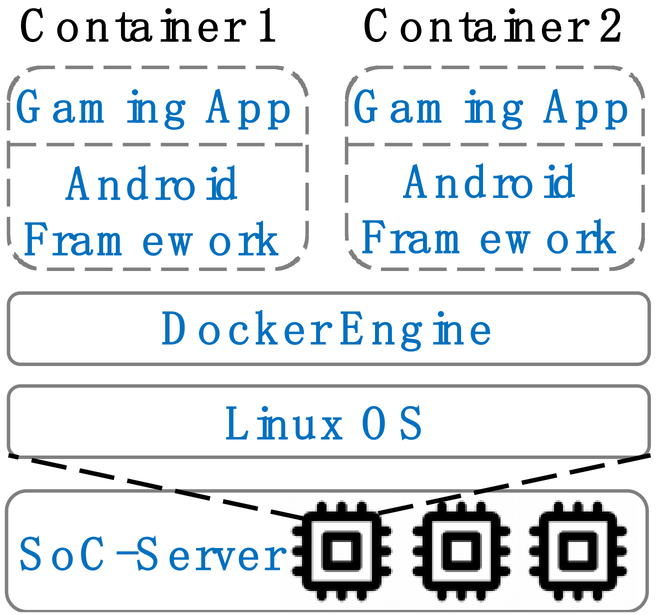
(b) The overview of a physical server.

(c) The internal PCB board with 5 SoCs.

Potential Killer Apps

- Mobile cloud gaming
  - De facto app served by SoC-Clusters
  - Business success: Genshin Impact
  - Running natively on mobile SoCs with Android container

Cloud Genshin Impact



Cloud Gaming Software Arch

Potential Killer Apps

- Mobile cloud gaming
  - De facto app served by SoC-Clusters
  - Business success: Genshin Impact
  - Running natively on mobile SoCs with Android container

- Challenges
  - Performance isolation for multi-game parallelism
  - Resource-intensive games on out-of-date SoCs

Potential Killer Apps

- Live video transcoding
  - Dominant use case of public edge platforms (e.g., video conference, live streaming)
  - SoC-Cluster is good at this with its *Low-power CPUs* and *hardware codecs*

Potential Killer Apps

- Live video transcoding
  - Dominant use case of public edge platforms (e.g., video conference, live streaming)
  - SoC-Cluster is good at this with its *Low-power CPUs* and *hardware codecs*
- Challenges
  - FFmpeg on SoC CPUs works well, but doesn’t support encoding on hardware codec

	Decoder			Encoder		Other support		
	Internal	Standalone	Hardware output	Standalone	Hardware input	Filtering	Hardware context	Usable from ffmpeg CLI
AMF	N	N	N	Y	Y	N	Y	Y
NVENC/NVDEC/CUVID	N	Y	Y	Y	Y	Y	Y	Y
Direct3D 11	Y	-	Y	-	-	[illegible]	Y	Y
Direct3D 9 / DXVA2	Y	-	Y	-	-	N	Y	Y
libmfx	-	Y	Y	Y	Y	Y	Y	Y
MediaCodec	-	Y	Y	N	N	-	N	N
Media Foundation	-	N	N	N	N	N	N	N
MMAL	-	Y	Y	N	N	-	N	N
OpenCL	-	-	-	-	-	Y	Y	Y

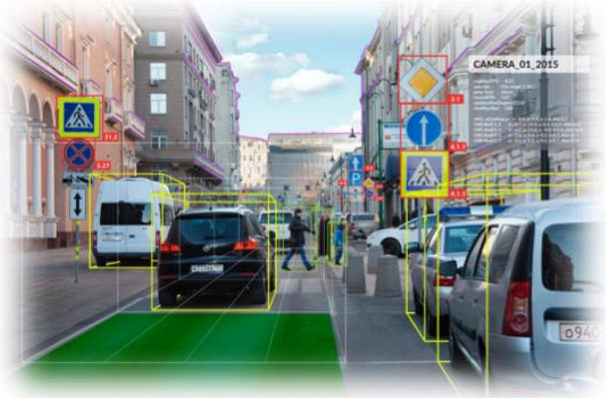
FFmpeg for Android only supports decoding but not encoding

Potential Killer Apps

- Live video transcoding
  - Dominant use case of public edge platforms (e.g., video conference, live streaming)
  - SoC-Cluster is good at this with its *Low-power CPUs* and *hardware codecs*
- Challenges
  - FFmpeg on SoC CPUs works well, but doesn't support encoding on hardware codec
  - HW-accel transcoding demand: LinkedIn LiTr for single video transcoding
  - Lack of unified task scheduling framework

Potential Killer Apps

- Deep learning serving
  - Use cases at the edge: AR/VR, autonomous driving, intelligent cameras
  - A good fit for SoC-Cluster: (1) energy-intensive (2) heterogeneous processors like GPU, DSP, NPU for acceleration
    - Mobile DL is blossoming!
  - Impressive energy-efficiency and comparable throughput (shown in later exp)



Autonomous driving

Speech Translation

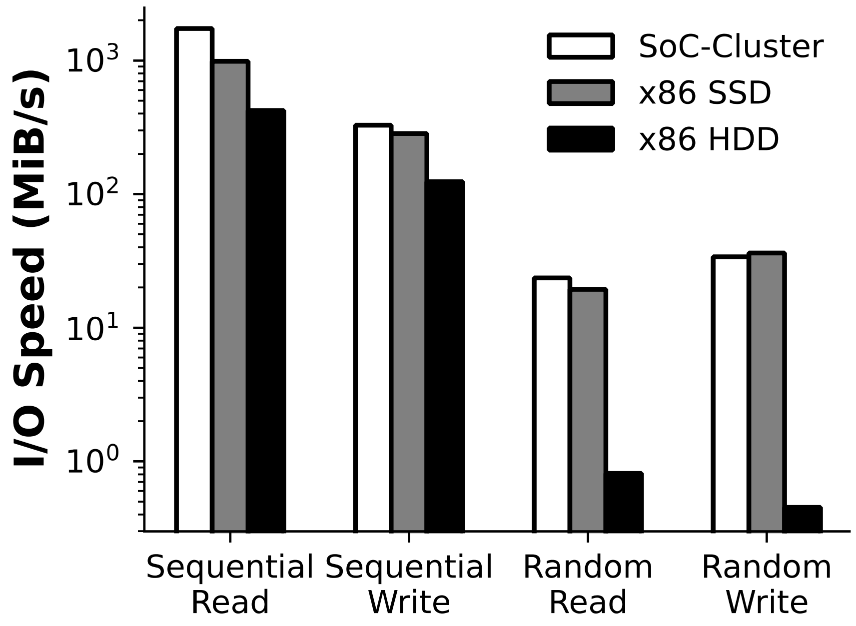
Intelligent Cameras

Potential Killer Apps

- Deep learning serving
  - Use cases at the edge: AR/VR, autonomous driving, intelligent cameras
  - A good fit for SoC-Cluster: (1) energy-intensive (2) heterogeneous processors like GPU, DSP, NPU for acceleration
    - Mobile DL is blossoming!
  - Impressive energy-efficiency and comparable throughput (shown in later exp)
- Challenges
  - High inference latency on large models: collaborative inference across SoCs for large DNN models (e.g., YOLOv5x, ResNet-152)

Potential Killer Apps

- Database systems
  - Basic building block of Internet services: Amazon Dynamo, Meta memcached, etc.
  - I/O intensive apps: massive parallelism, independent concurrent operations
  - Fast flash storage on each SoC!



Measured by fio

- **Each SoC: 256GB SK-Hynix flash storage**
  - **Sequential R/W: 1,733 and 328 MiB/s**
  - **Random R/W: 24 and 34 MiB/s**
- **Performance: comparable to an enterprise Samsung SSD, faster than a Seagate HDD on traditional edge servers.**
- **In total, 15.36 TB storage, 1GiB/s rand I/O.**

Potential Killer Apps

- Database systems
  - Basic building block of Internet services: Amazon Dynamo, Meta memcached, etc.
  - I/O intensive apps: massive parallelism, independent concurrent operations
  - Fast flash storage on each SoC!
- Challenges
  - Distribute data across SoCs to ensure I/O operations can be concurrently handled without congestion

Potential Killer Apps

- Database systems
  - Basic building block of Internet services: Amazon Dynamo, Meta memcached, etc.
  - I/O intensive apps: massive parallelism, independent concurrent operations
  - Fast flash storage on each SoC!
- Challenges
  - Distribute data across SoCs to ensure I/O operations can be concurrently handled without congestion

FAWN: A Fast Array of Wimpy Nodes

David G. Andersen[1], Jason Franklin[1], Michael Kaminsky[2], Amar Phanishayee[1], Lawrence Tan[1], Vijay Vasudevan[1]

[1]*Carnegie Mellon University,* [2]*Intel Labs*

Potential Killer Apps

- Stream processing
  - Massive data generated by IoT devices at the edge
  - Fit for SoC-Cluster: multiple CPU cores (60 * 8) and enough memory bandwidth (60 * LPDDR5 DRAM)

Potential Killer Apps

- Stream processing
  - Massive data generated by IoT devices at the edge
  - Fit for SoC-Cluster: multiple CPU cores (60 * 8) and enough memory bandwidth (60 * LPDDR5 DRAM)

- Mobile-computation offloading
  - Run mobile-native software seamlessly
  - Offloading *hot spots* code regions
  - Critical challenge: low-latency state synchronization

Case studies

- Live video transcoding
  - Software: FFmpeg & LiTr[1]
  - Datasets: 3 videos picked from vbench[2]
  - Metrics: throughput, energy efficiency, video quality
- Deep learning serving
  - Software: TVM@Intel CPU; TensorRT@NVIDIA GPU; TFLite@SoC
  - Model: ResNet-50 (FP32/INT8)
  - Metrics: latency, throughput, energy efficiency
- Alternative hardware (a traditional edge server)
  - A 40-core Intel Xeon Gold 5218R processor
  - 8 * NVIDIA A40 GPU

Live video transcoding

Video	Hardware	Throughput (# of streams)	Energy (frames/J)	PSNR (db)
V1-desktop Bitrate: 180 Kbps	Intel CPU (4-core container)	31	23	31.08
	NVIDIA A40	37	13	34.11
	SoC CPU	15	59	31.21
	SoC Codec	16	125	29.27
V2-game3 Bitrate: 5.6 Mbps	Intel CPU (4-core container)	8	11	39.69
	NVIDIA A40	18	12	40.73
	SoC CPU	4	32	40.37
	SoC Codec	12	167	34.72
V3-chicken Bitrate: 49 Mbps	Intel CPU (4-core container)	2	2	38.71
	NVIDIA A40	6	2	42.54
	SoC CPU	1	5	38.80
	SoC Codec	2	26	38.28

TABLE II

LIVE VIDEO TRANSCODING PERFORMANCE OF SOC-CLUSTER AND CONVENTIONAL SERVERS. VIDEOS ARE PICKED FROM A CLOUD VIDEO TRANSCODING BENCHMARK [56].

- **SoC hardware codec improves throughput up to 3x (compared to SoC CPU).**
- **An SoC-Cluster can transcode180–1,860 streams collectively.**
  - **40-core Intel CPU: 20-310 streams**
  - **Equals to 30-53 A40 GPUs**

Substantially higher throughput even on SoC CPU.

Live video transcoding

北京邮電大學
Beijing University of Posts and Telecommunications

Video	Hardware	Throughput (# of streams)	Energy (frames/J)	PSNR (db)
V1-desktop Bitrate: 180 Kbps	Intel CPU (4-core container)	31	23	31.08
	NVIDIA A40	37	13	34.11
	SoC CPU	15	59	31.21
	SoC Codec	16	125	29.27
V2-game3 Bitrate: 5.6 Mbps	Intel CPU (4-core container)	8	11	39.69
	NVIDIA A40	18	12	40.73
	SoC CPU	4	32	40.37
	SoC Codec	12	167	34.72
V3-chicken Bitrate: 49 Mbps	Intel CPU (4-core container)	2	2	38.71
	NVIDIA A40	6	2	42.54
	SoC CPU	1	5	38.80
	SoC Codec	2	26	38.28

TABLE II

LIVE VIDEO TRANSCODING PERFORMANCE OF SOC-CLUSTER AND CONVENTIONAL SERVERS. VIDEOS ARE PICKED FROM A CLOUD VIDEO TRANSCODING BENCHMARK [56].

- **SoC Codec can transcode 26–167 frames per Joule, up to 15.18× higher than the Intel CPU and up to 13.92× higher than the NVIDIA A40 GPU.**

SoC CPU/Codec both deliver higher energy efficiency!

Live video transcoding

Video	Hardware	Throughput (# of streams)	Energy (frames/J)	PSNR (db)
V1-desktop Bitrate: 180 Kbps	Intel CPU (4-core container)	31	23	31.08
	NVIDIA A40	37	13	34.11
	SoC CPU	15	59	31.21
	SoC Codec	16	125	29.27
V2-game3 Bitrate: 5.6 Mbps	Intel CPU (4-core container)	8	11	39.69
	NVIDIA A40	18	12	40.73
	SoC CPU	4	32	40.37
	SoC Codec	12	167	34.72
V3-chicken Bitrate: 49 Mbps	Intel CPU (4-core container)	2	2	38.71
	NVIDIA A40	6	2	42.54
	SoC CPU	1	5	38.80
	SoC Codec	2	26	38.28

TABLE II

LIVE VIDEO TRANSCODING PERFORMANCE OF SOC-CLUSTER AND CONVENTIONAL SERVERS. VIDEOS ARE PICKED FROM A CLOUD VIDEO TRANSCODING BENCHMARK [56].

- **SoC CPU (SW encoder): almost the same quality as Intel CPU/NVIDIA GPU.**
- **SoC Codec (HW encoder): slightly poorer quality than others.**
  - **Mainly due to the loose quality/bitrate requirements of mobile encoders inherently.**

SoC CPU is more suitable for quality-sensitive apps.

Deep learning serving

Model	Hardware	Latency (ms)	Throughput (frames/s)	Energy (frames/J)
ResNet-50 (FP32)	Intel CPU (4 cores)	12.47	80	2.6
	NVIDIA A40 (BS=1)	2.18	459	2.8
	NVIDIA A40 (BS=64)	23.45	2,729	10.2
	SoC CPU (4 big cores)	77.60	13	2.1
	SoC GPU	32.70	31	18.2
ResNet-50 (INT8)	NVIDIA A40 (BS=1)	0.45	2,202	18.6
	NVIDIA A40 (BS=64)	7.51	8,526	31.3
	SoC DSP	8.80	114	71.4

TABLE III

DL SERVING PERFORMANCE OF SOC-CLUSTER AND CONVENTIONAL EDGE SERVERS. DEFAULT BATCH SIZE (BS) IS 1.

- **For FP32 model: an SoC-Cluster delivers a max throughput at 2,640 FPS.**
  - **Equals to 132-core Intel CPU.**
  - **Equals to ~1 NVIDIA A40 GPU.**
- **For INT8 model: an SoC-Cluster delivers a max throughput at 6,840 FPS.**
  - **Close to a NVIDIA A40 GPU.**

Higher throughput than CPU servers;
Slightly lower throughput than GPU servers.

Deep learning serving

Model	Hardware	Latency (ms)	Throughput (frames/s)	Energy (frames/J)
ResNet-50 (FP32)	Intel CPU (4 cores)	12.47	80	2.6
	NVIDIA A40 (BS=1)	2.18	459	2.8
	NVIDIA A40 (BS=64)	23.45	2,729	10.2
	SoC CPU (4 big cores)	77.60	13	2.1
	SoC GPU	32.70	31	18.2
ResNet-50 (INT8)	NVIDIA A40 (BS=1)	0.45	2,202	18.6
	NVIDIA A40 (BS=64)	7.51	8,526	31.3
	SoC DSP	8.80	114	71.4

TABLE III

DL SERVING PERFORMANCE OF SOC-CLUSTER AND CONVENTIONAL EDGE SERVERS. DEFAULT BATCH SIZE (BS) IS 1.

- **SoC GPU/DSP provide higher energy efficiency than the traditional edge server.**
  - **SoC GPU is 7x/1.8x energy efficient than Intel CPU/NVIDIA A40 GPU.**
  - **SoC DSP shows higher energy efficiency than SoC GPU.**

Deep learning serving

Model	Hardware	Latency (ms)	Throughput (frames/s)	Energy (frames/J)
ResNet-50 (FP32)	Intel CPU (4 cores)	12.47	80	2.6
	NVIDIA A40 (BS=1)	2.18	459	2.8
	NVIDIA A40 (BS=64)	23.45	2,729	10.2
	SoC CPU (4 big cores)	77.60	13	2.1
	SoC GPU	32.70	31	18.2
ResNet-50 (INT8)	NVIDIA A40 (BS=1)	0.45	2,202	18.6
	NVIDIA A40 (BS=64)	7.51	8,526	31.3
	SoC DSP	8.80	114	71.4

TABLE III

DL SERVING PERFORMANCE OF SOC-CLUSTER AND CONVENTIONAL EDGE SERVERS. DEFAULT BATCH SIZE (BS) IS 1.

- **SoC-Cluster is good at scaling its workloads with low energy consumption.**
  - **NVIDIA GPU: BS (64 to 1), EF (10.2 to 2.8).**
  - **SoC-Cluster: each SoC works with BS=1, low power.**
  - **Idle SoCs can also be turned off!**

Deep learning serving

Model	Hardware	Latency (ms)	Throughput (frames/s)	Energy (frames/J)
ResNet-50 (FP32)	Intel CPU (4 cores)	12.47	80	2.6
	NVIDIA A40 (BS=1)	2.18	459	2.8
	NVIDIA A40 (BS=64)	23.45	2,729	10.2
	SoC CPU (4 big cores)	77.60	13	2.1
	SoC GPU	32.70	31	18.2
ResNet-50 (INT8)	NVIDIA A40 (BS=1)	0.45	2,202	18.6
	NVIDIA A40 (BS=64)	7.51	8,526	31.3
	SoC DSP	8.80	114	71.4

TABLE III
DL SERVING PERFORMANCE OF SOC-CLUSTER AND CONVENTIONAL EDGE SERVERS. DEFAULT BATCH SIZE (BS) IS 1.

- **SoC GPU/DSP deliver much lower latency than its CPU.**
  - **8.8ms on SoC DSP is eligible for most edge apps!**
- **NVIDIA GPU delivers much lower latency with a small batch size, but a higher energy cost.**
- **A single SoC is not likely to achieve satisfactory latency on large DNNs (e.g., YOLOv5x, BERT).**

Takeaways

- Time to reconsider the edge server design
  - Why inherit the legacy from clouds
- An extreme design: SoC-Cluster
  - Massive, low-power, sub-10 nm chips.
  - Each SoC is heterogeneous itself (with GPU/NPU).
  - Commercial success in mobile cloud gaming services.
- A set of experiments that demonstrates the pros/cons of SoC-Cluster over traditional servers.

Open to any discussion or debate!

mwx@bupt.edu.cn