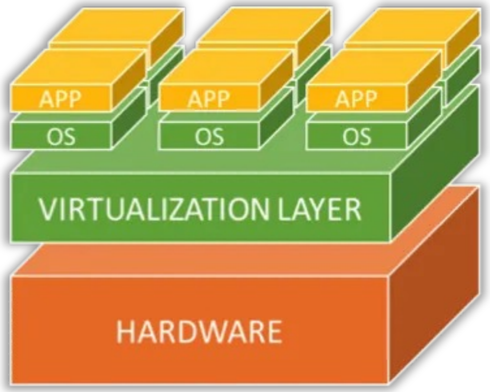
More is Different
Prototyping and Analyzing a New Form of Edge Server with Massive Mobile SoCs

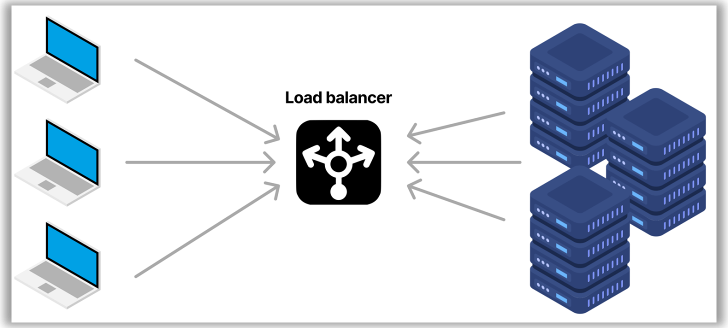
Li Zhang[1], Zhe Fu[2], Boqing Shi[1], Xiang Li[1], Rujin Lai[3], Chenyang Yang[3], Ao Zhou[1], Xiao Ma[1], Shangguang Wang[1], Mengwei Xu[1]

[1]*Beijing University of Posts and Telecommunications (BUPT)*
[2]*Tsinghua University*, [3]*vclusters*

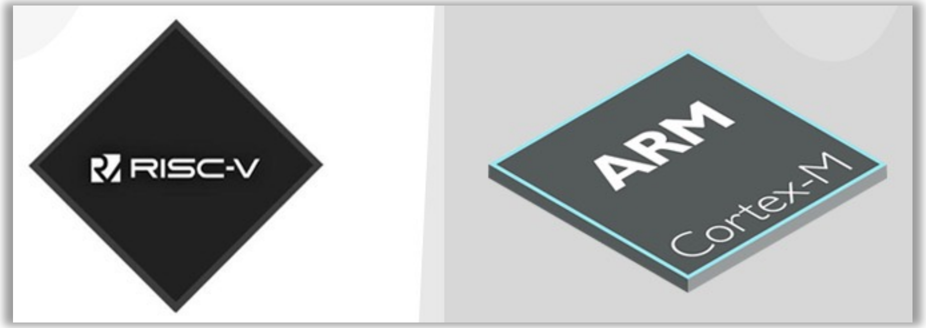
Power Consumption of Datacenters

- High power consumption: A historical and persistent issue
- Workloads: Large language model training/inference, big data analytics, video streaming, etc.
- How to build energy-efficient datacenters?
  - Software optimizations: Resource virtualization, load balancing, …
  - Hardware optimizations: Use RISC architecture, lower process nodes, …





Software-level Optimizations

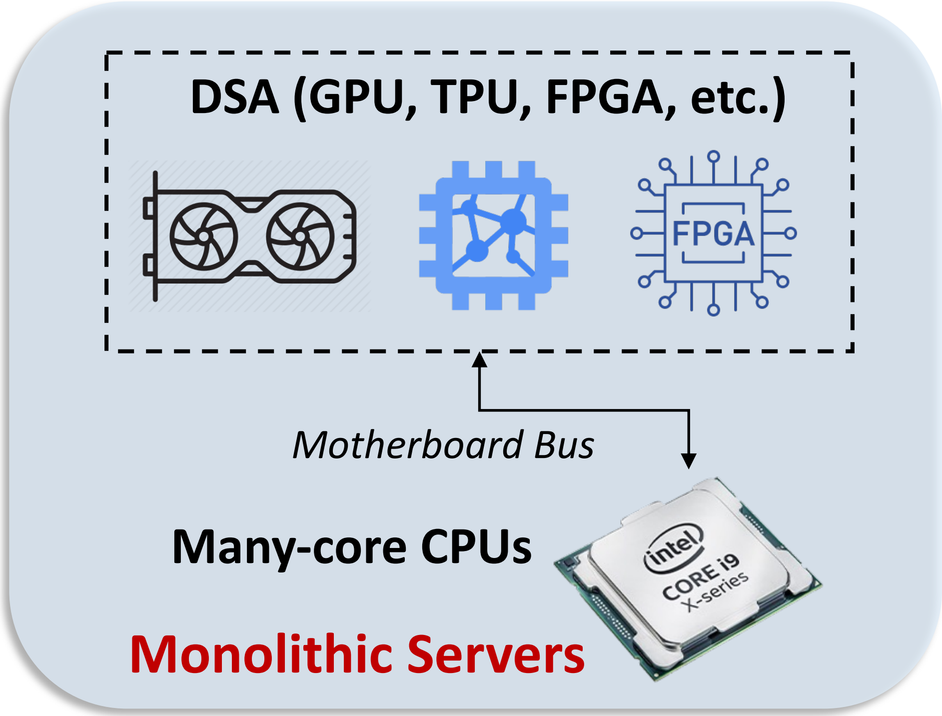


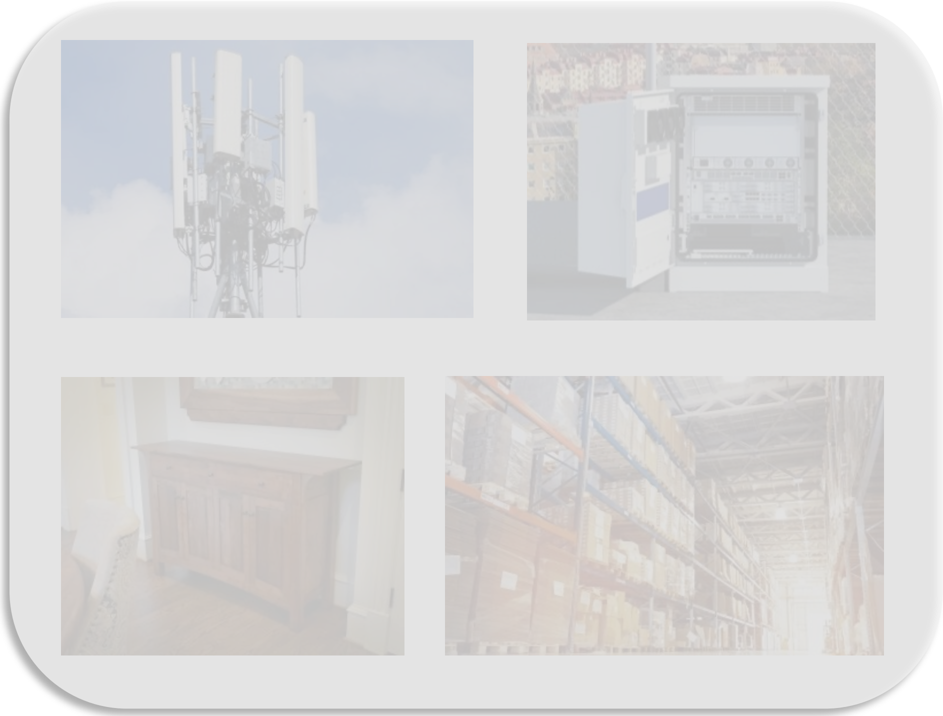
Hardware-level Optimizations

Cloud vs. Edge: Key Factors

Cherry-picked and Large	Location and Spaces	Near-population and Limited
Abundant and Cheap	Power Supply	Constrained and Expensive
Powerful and Mature	Colling Facility	Wimpy or Even Missed
Various Types, Stable/Predictable	Workloads	Specific and Highly Variational

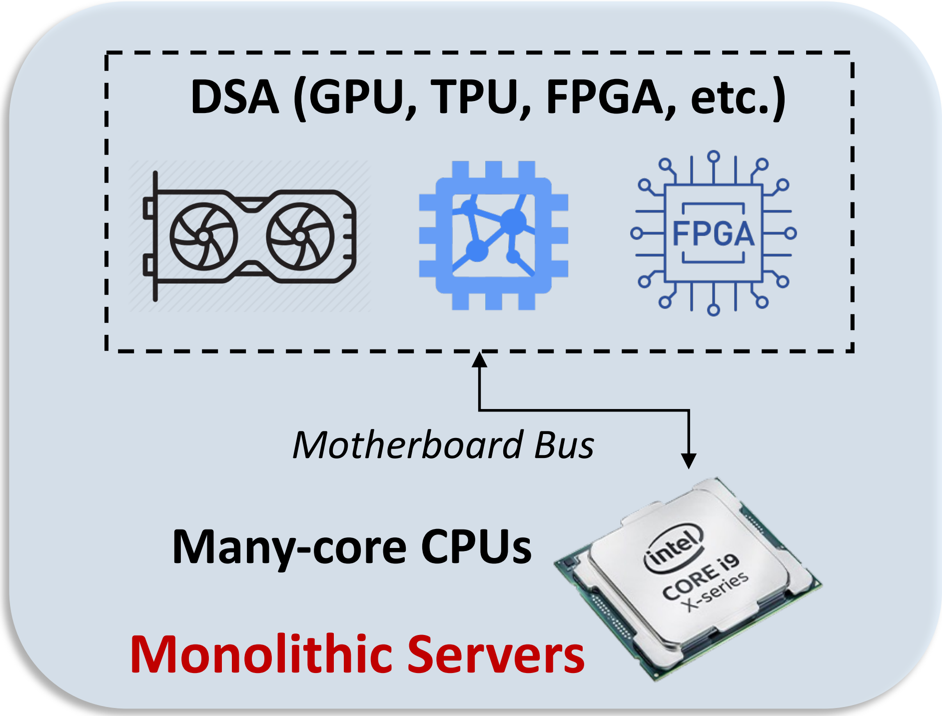
Cloud vs. Edge: Hardware Selection

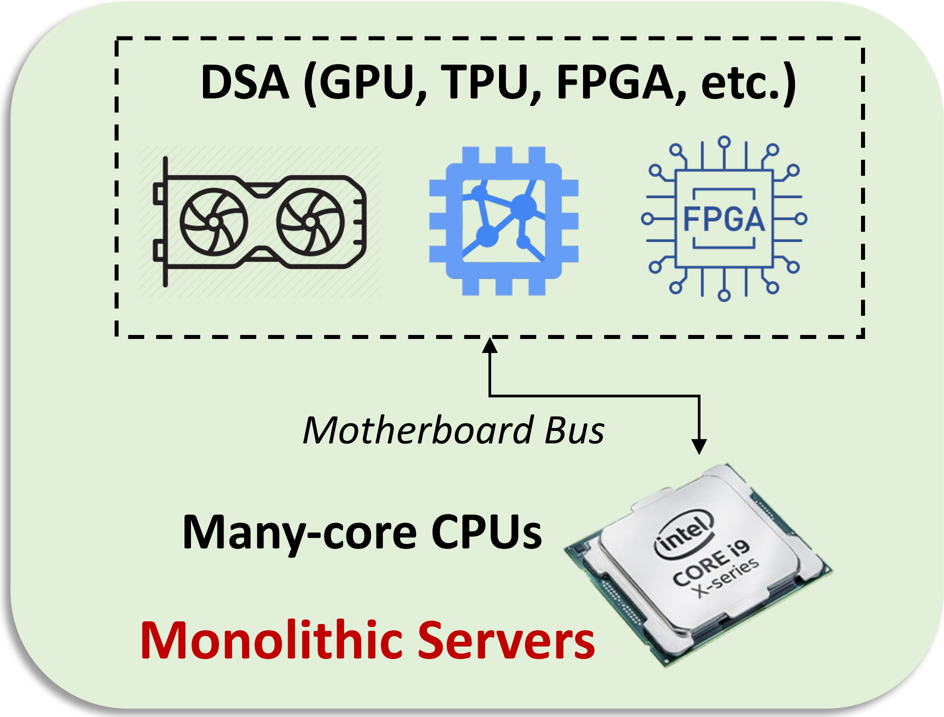




Cherry-picked and Large	<u>Location and Spaces</u>	Near-population and Limited
Abundant and Cheap	<u>Power Supply</u>	Constrained and Expensive
Powerful and Mature	<u>Colling Facility</u>	Wimpy or Even Missed
Various Types, Stable/Predictable	<u>Workloads</u>	Specific and Highly Variational

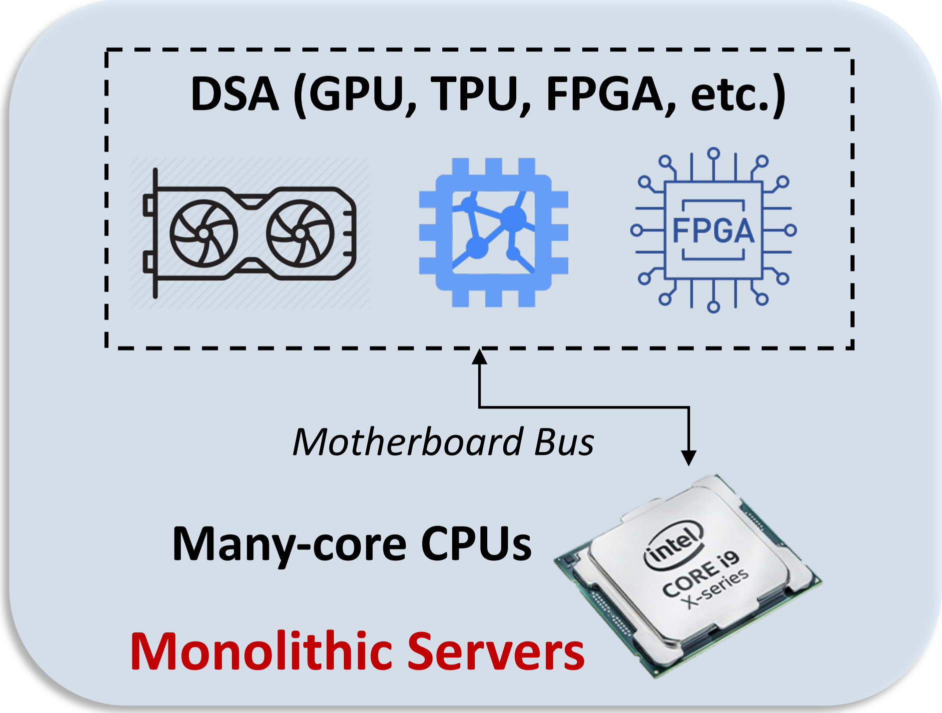
Cloud vs. Edge: Hardware Selection





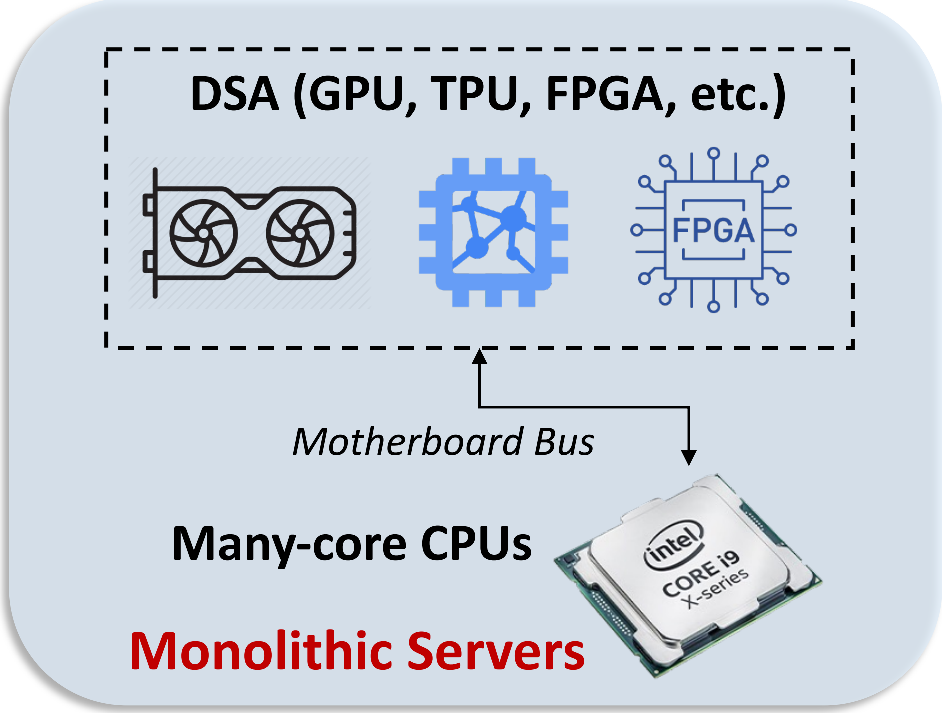
Cherry-picked and Large	Location and Spaces	Near-population and Limited
Abundant and Cheap	Power Supply	Constrained and Expensive
Powerful and Mature	Colling Facility	Wimpy or Even Missed
Various Types, Stable/Predictable	Workloads	Specific and Highly Variational

Cloud vs. Edge: Hardware Selection



Cherry-picked and Large	Location and Spaces	Near-population and Limited
Abundant and Cheap	Power Supply	Constrained and Expensive
Powerful and Mature	Colling Facility	Wimpy or Even Missed
Various Types, Stable/Predictable	Workloads	Specific and Highly Variational

Cloud vs. Edge: Hardware Selection



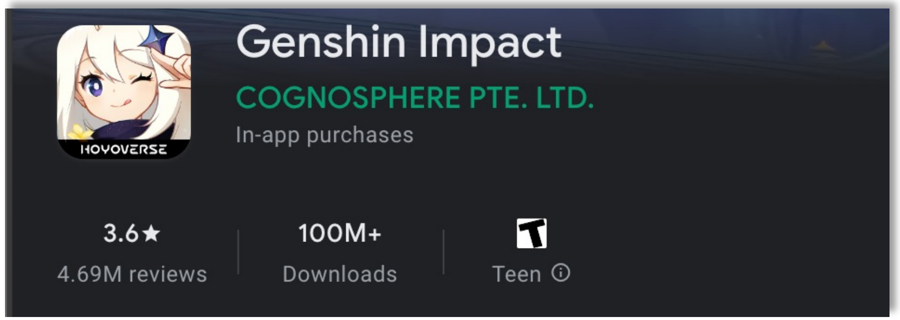
Smartphones!

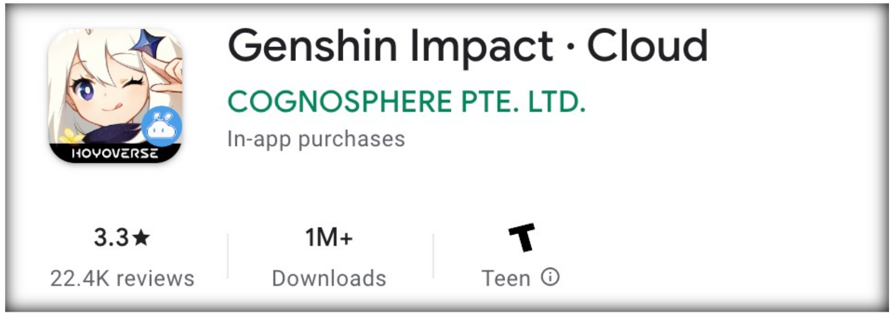
Intrinsic benefits of smartphones at the edge

- ❑ Higher energy efficiency of mobile processors compared to traditional datacenter servers
- ❑ Heterogeneous co-processors like mobile GPU, NPU, video codec, etc.
- ❑ Ability to run mobile operating systems and apps

Killer Workload: Mobile Cloud Gaming

- Mobile cloud gaming services: Enable wimpy mobile devices to run immersive, resource-consuming (computing and disk) mobile games released in recent years.
- Business success: Genshin Impact gains
  - > 5B USD income dated to Feb. 2024.
  - > 1M downloads of its cloud gaming version dated to July 2024.





- Underlying rationale: Mobile games are optimized for mobile platforms/processors

Killer Workload: Mobile Cloud Gaming

High-density Mobile Cloud Gaming on Edge SoC Clusters

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

Abstract

System-on-Chip (SoC) Clusters, i.e., servers consisting of many stacked mobile SoCs, have emerged as a popular platform for serving mobile cloud gaming. Sharing the underlying hardware and OS, these SoC Clusters enable native mobile games to be executed and rendered efficiently without modification. However, the number of deployed game sessions is limited due to conservative deployment strategies and high GPU utilization in current game offloading methods. To address these challenges, we introduce *SFG*, the first system that enables high-density mobile cloud gaming on SoC Clusters with two novel techniques: (1) It employs a resource-efficient game partitioning and cross-SoC offloading design that maximally preserves GPU optimization intents in the standard graphics rendering pipeline; (2) It proposes

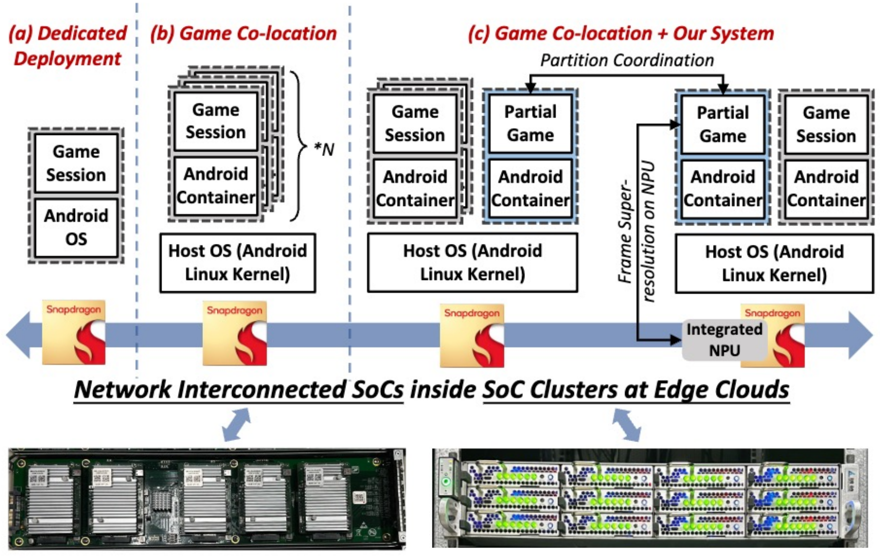
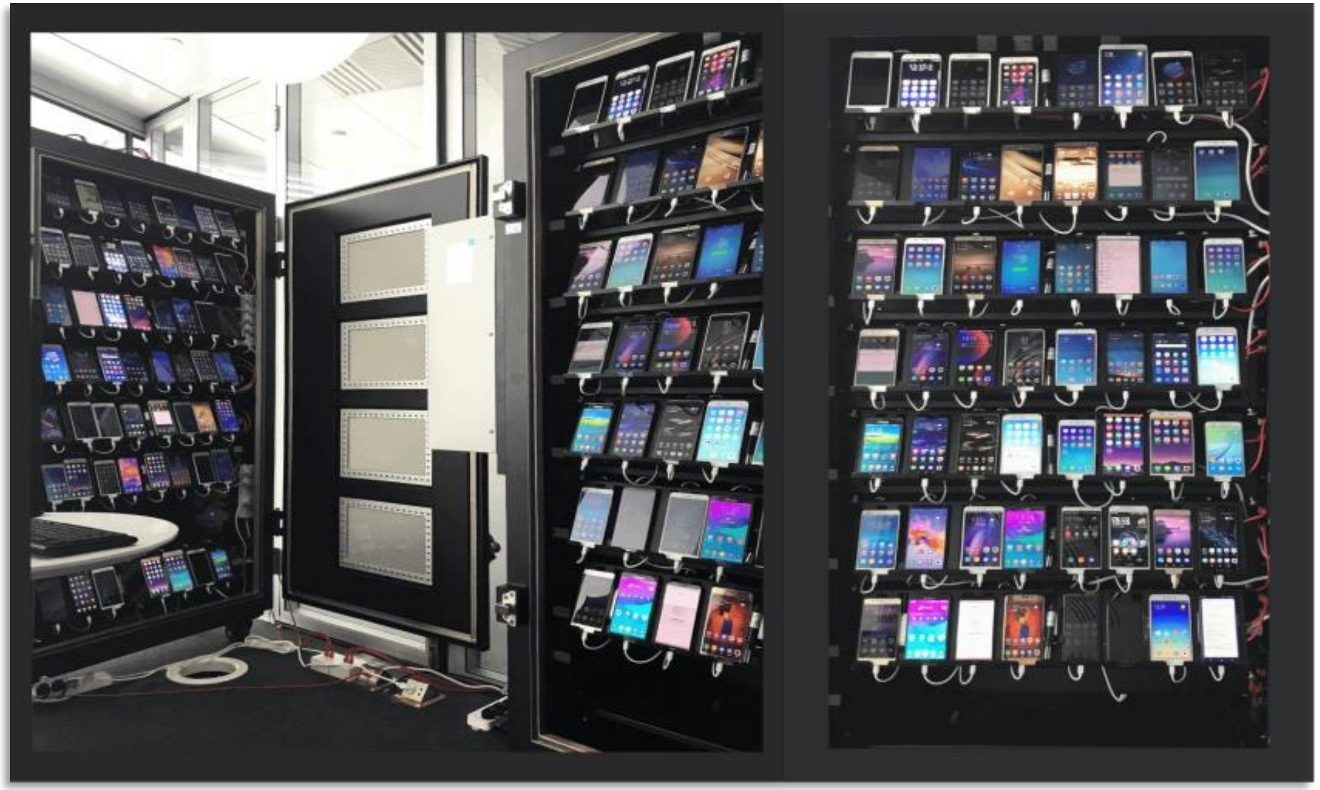


Figure 1: Hardware/software architecture and different deployment strategies of mobile gaming on SoC Clusters.

Massive Smartphones in the Cloud

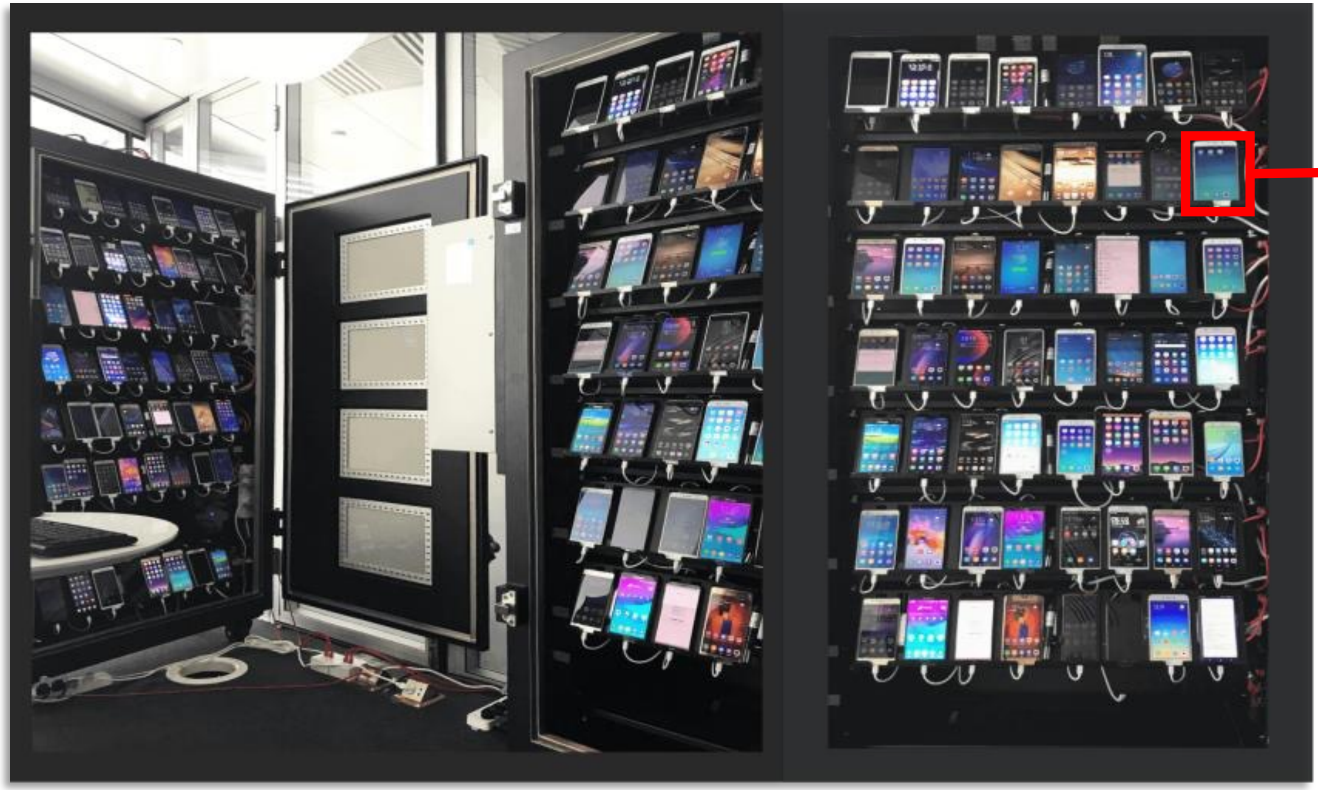


Physical Smartphone Farms

AWS Device Farm, Google Firebase Test Lab, Douyin Device Farm[1]

[1] [MobiCom'23] Hao Lin et al. Virtual Device Farms for Mobile App Testing at Scale: A Pursuit for Fidelity, Efficiency, and Accessibility

Massive Mobile SoCs at the Edge

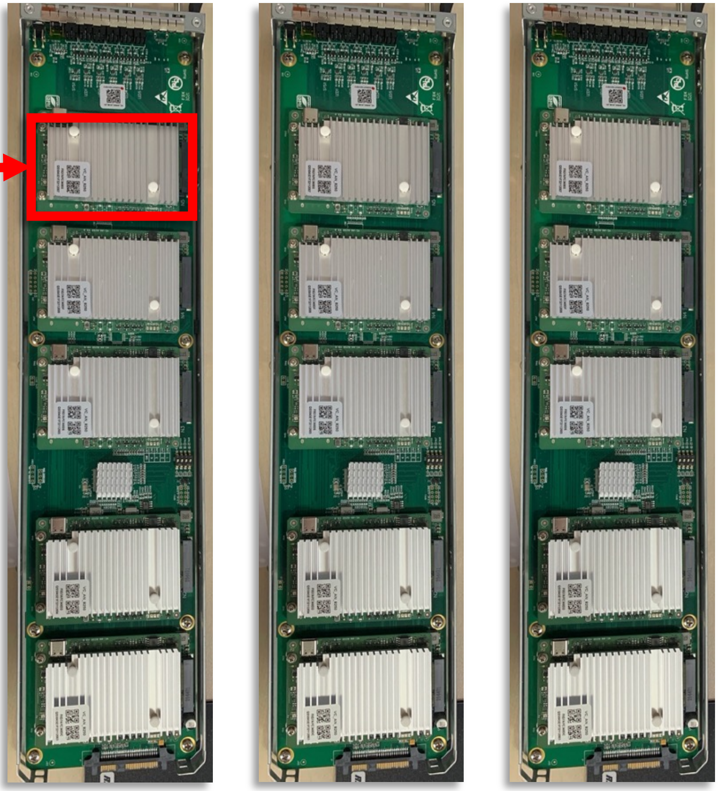


✗ Screen

✗ Battery

✅ Direct Power Supply

✅ Ethernet



Physical Smartphone Farms → **Massive Individual Mobile SoCs**

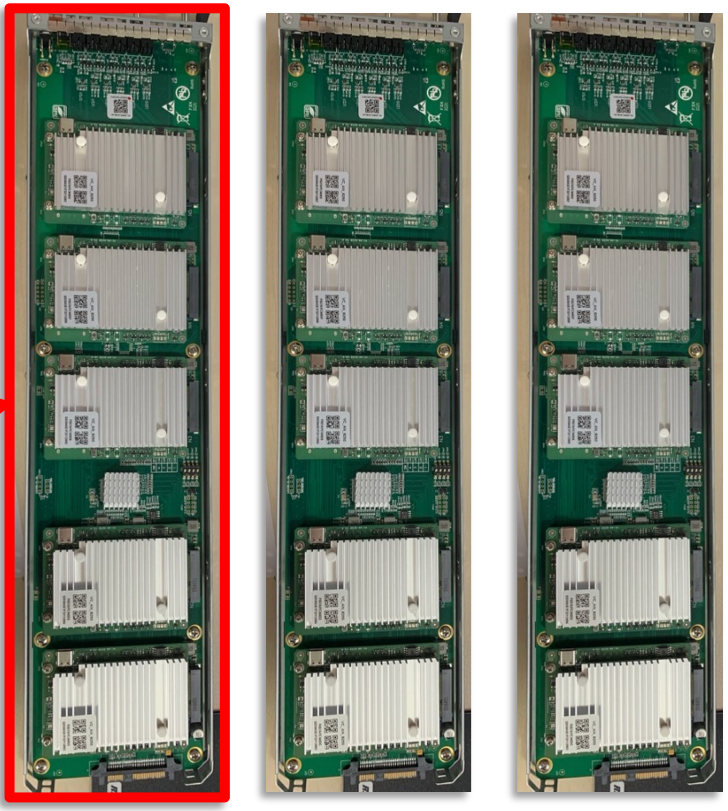
AWS Device Farm, Google Firebase Test Lab, Douyin Device Farm[1]

Stability & Higher Density & Higher Energy Efficiency

[1] [MobiCom'23] Hao Lin et al. Virtual Device Farms for Mobile App Testing at Scale: A Pursuit for Fidelity, Efficiency, and Accessibility

A Close Look at an SoC Cluster

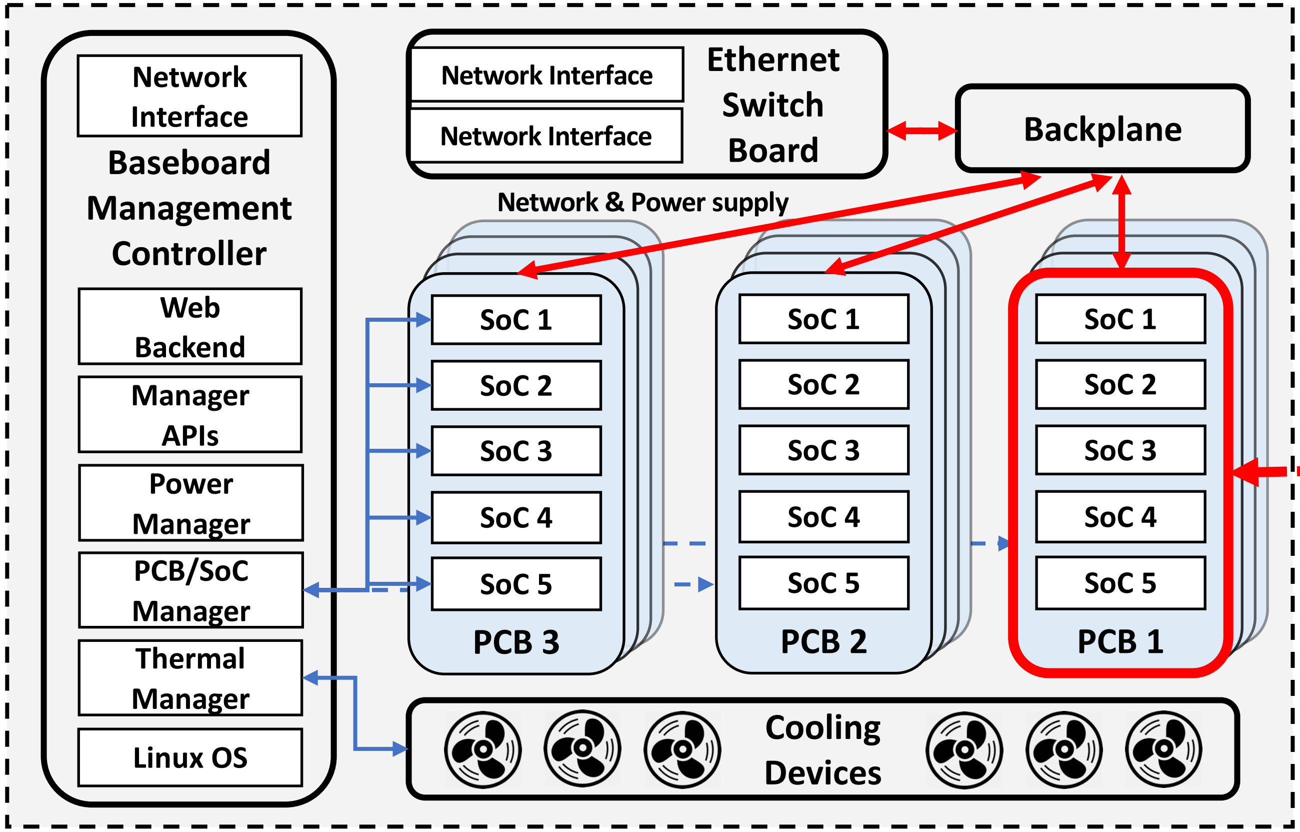
A commercial SoC Cluster



Massive Individual Mobile SoCs

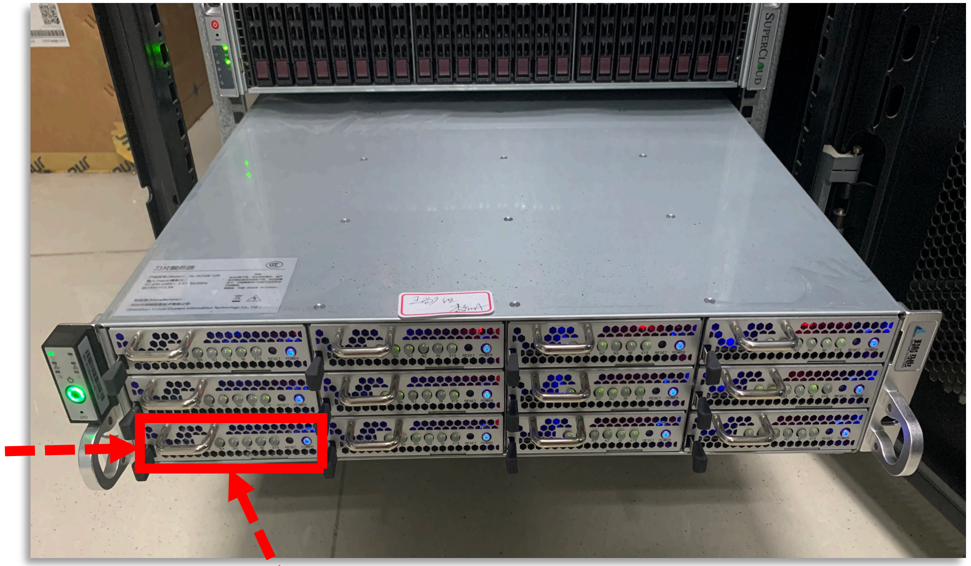
- ❑ In-the-wild deployment in edge clouds
- ❑ Support mobile cloud gaming, cloud phone services

A Close Look at an SoC Cluster



The Conceptual Architecture of an SoC Cluster

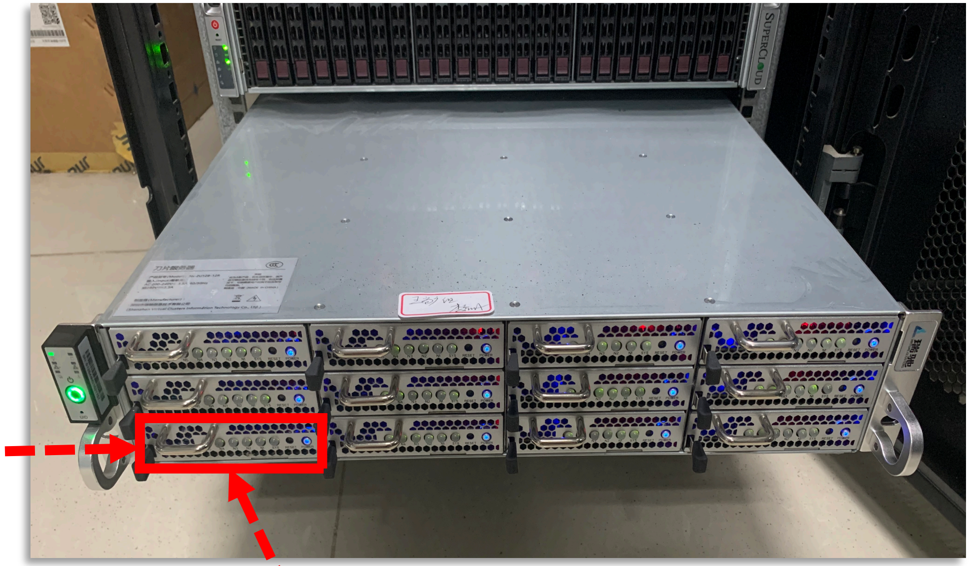
A Physcal SoC Cluster

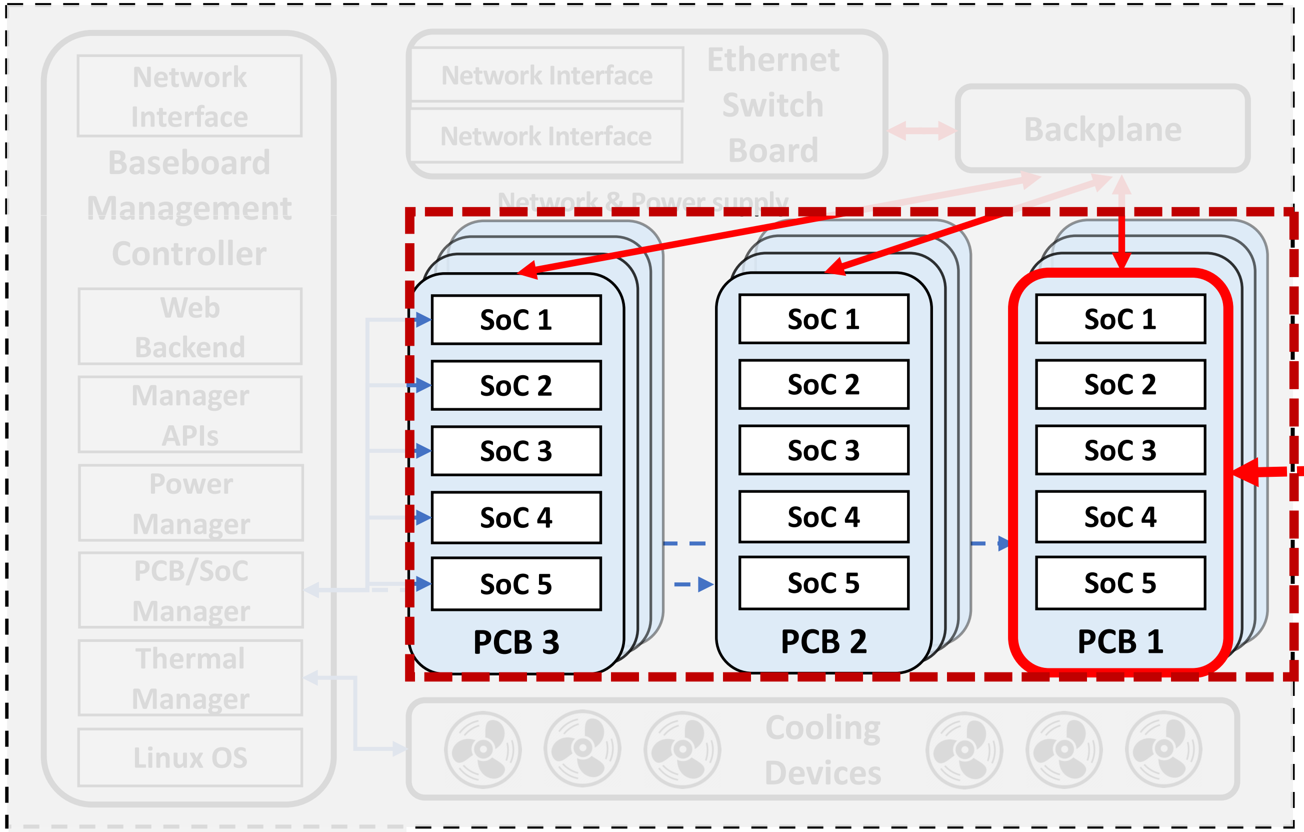


A Internal PCB board with 5 SoCs

A Close Look at an SoC Cluster

A Physcal SoC Cluster



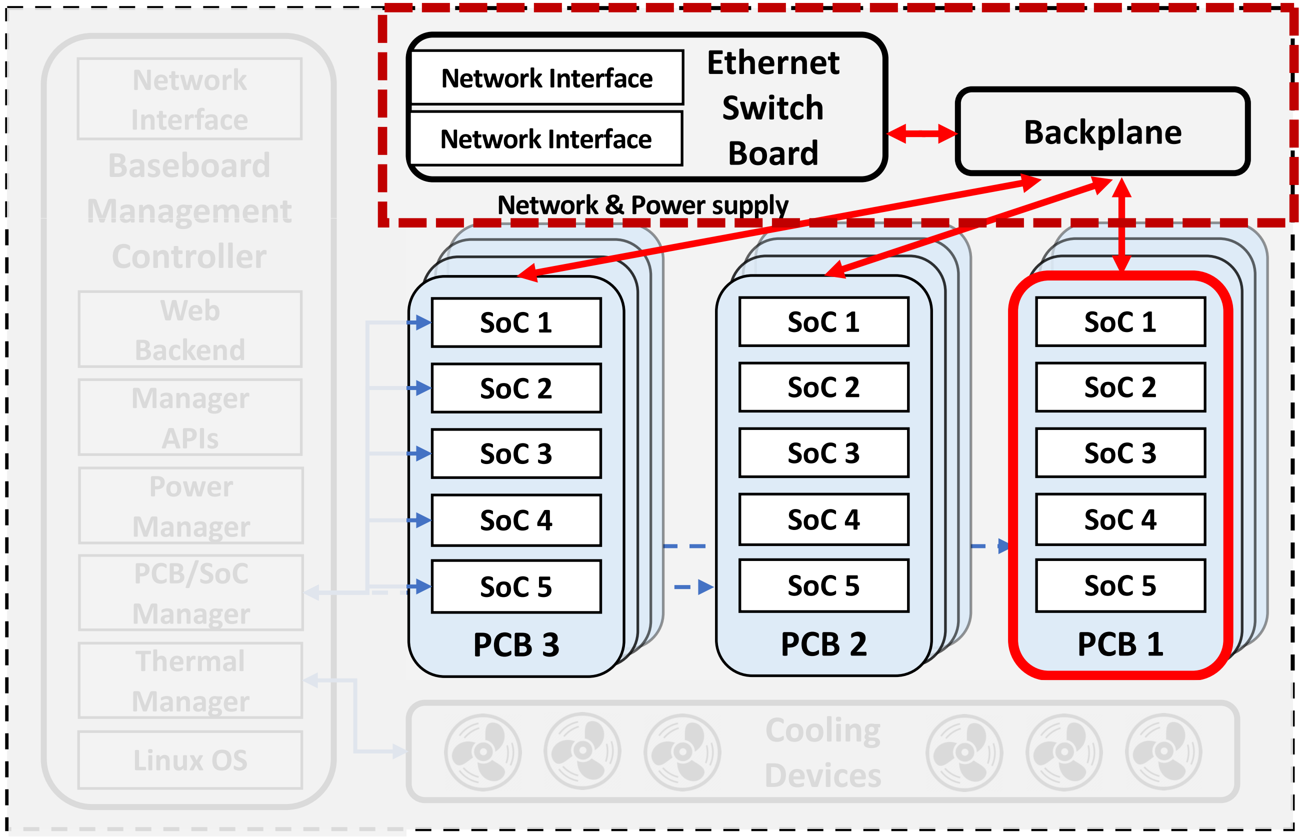


The Conceptual Architecture of an SoC Cluster

A Internal PCB board with 5 SoCs

- Computing units: Every 5 mobile SoCs are integrated into one printable circuit board (PCB). (60 SoCs in total)

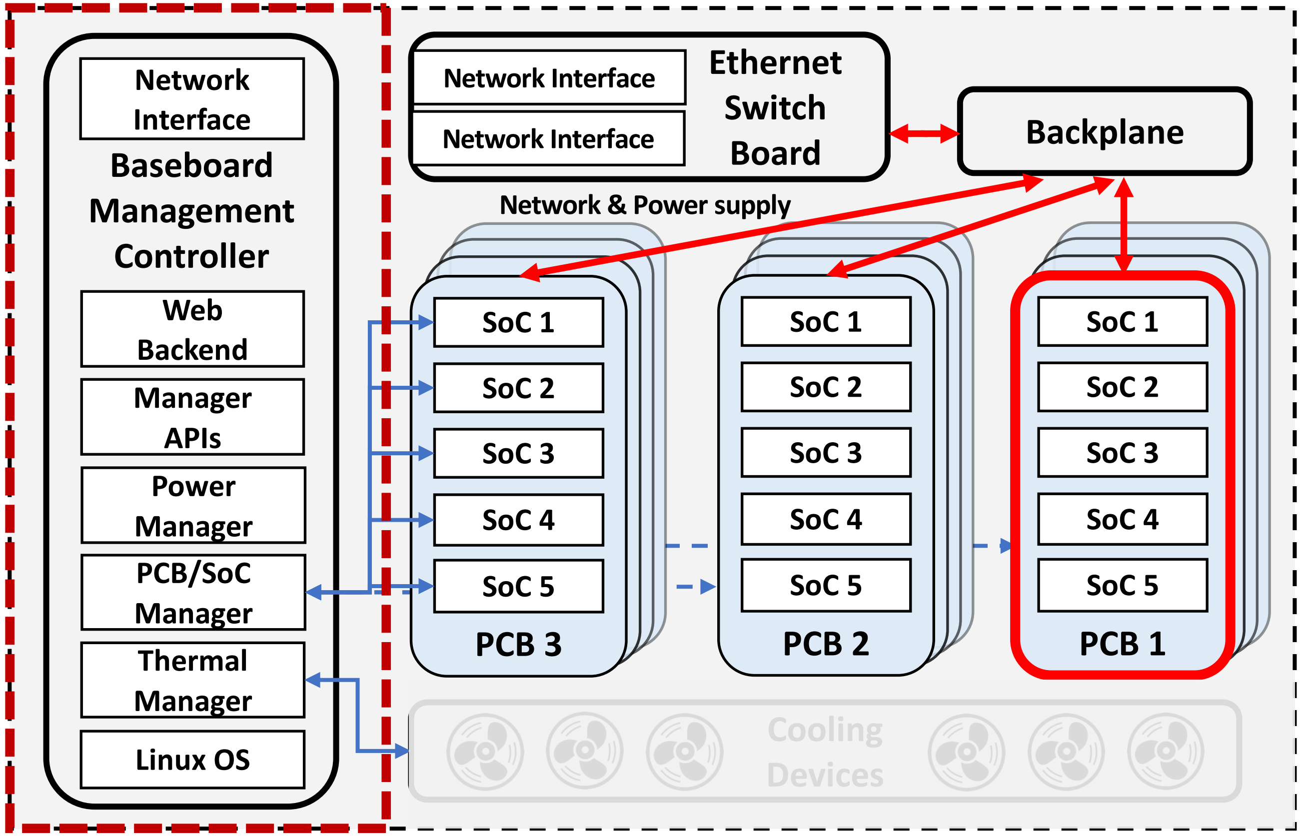
A Close Look at an SoC Cluster



The Conceptual Architecture of an SoC Cluster

- ❑ Computing units: Every 5 mobile SoCs are integrated into one printable circuit board (PCB). (60 SoCs in total)
- ❑ Networking: One backplane and one Ethernet Switch Board (20 Gbps); 2-layer Ethernet networking and power supply

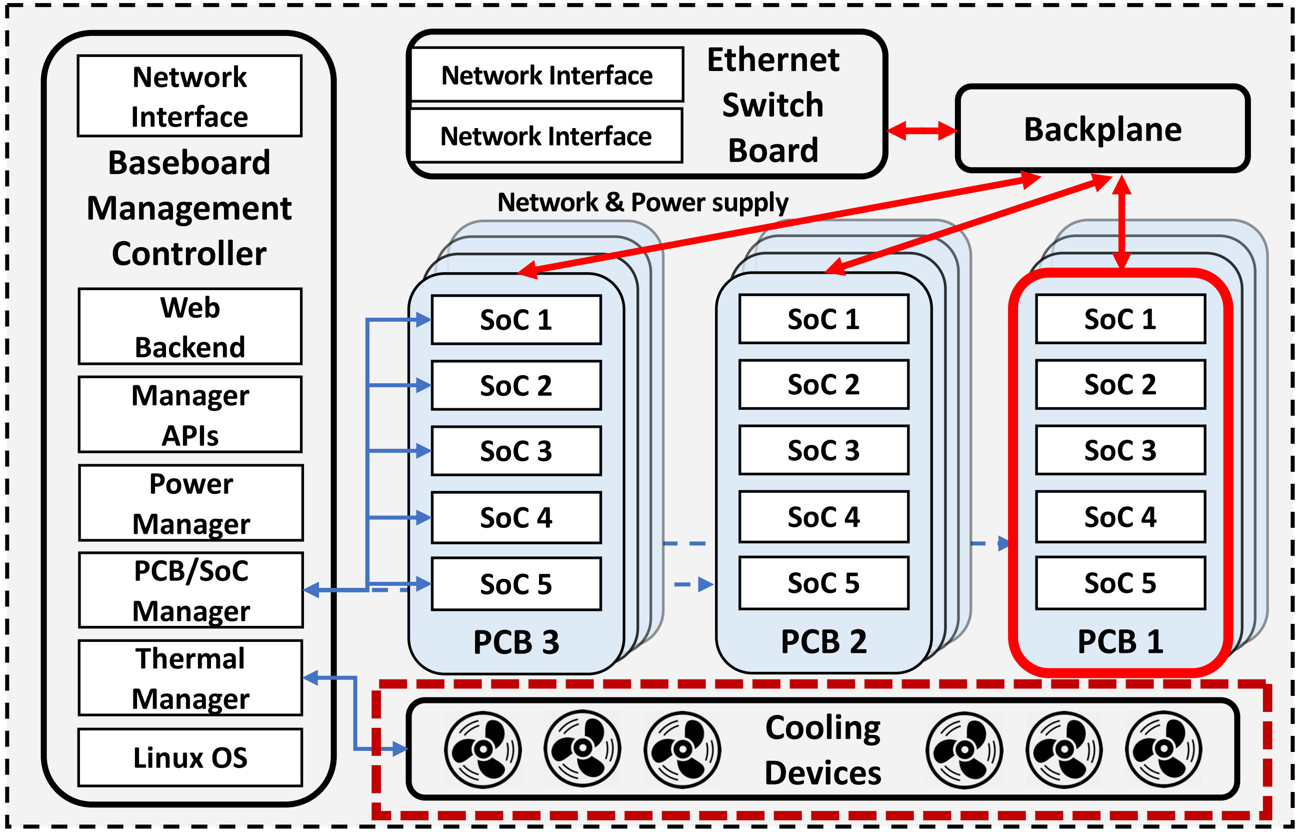
A Close Look at an SoC Cluster



The Conceptual Architecture of an SoC Cluster

- ❑ Computing units: Every 5 mobile SoCs are integrated into one printable circuit board (PCB). (60 SoCs in total)
- ❑ Networking: One backplane and one Ethernet Switch Board (20 Gbps); 2-layer Ethernet networking and power supply
- ❑ Server management: One baseboard management controller (BMC); SoC/PCB controller, thermal manager, etc.

A Close Look at an SoC Cluster

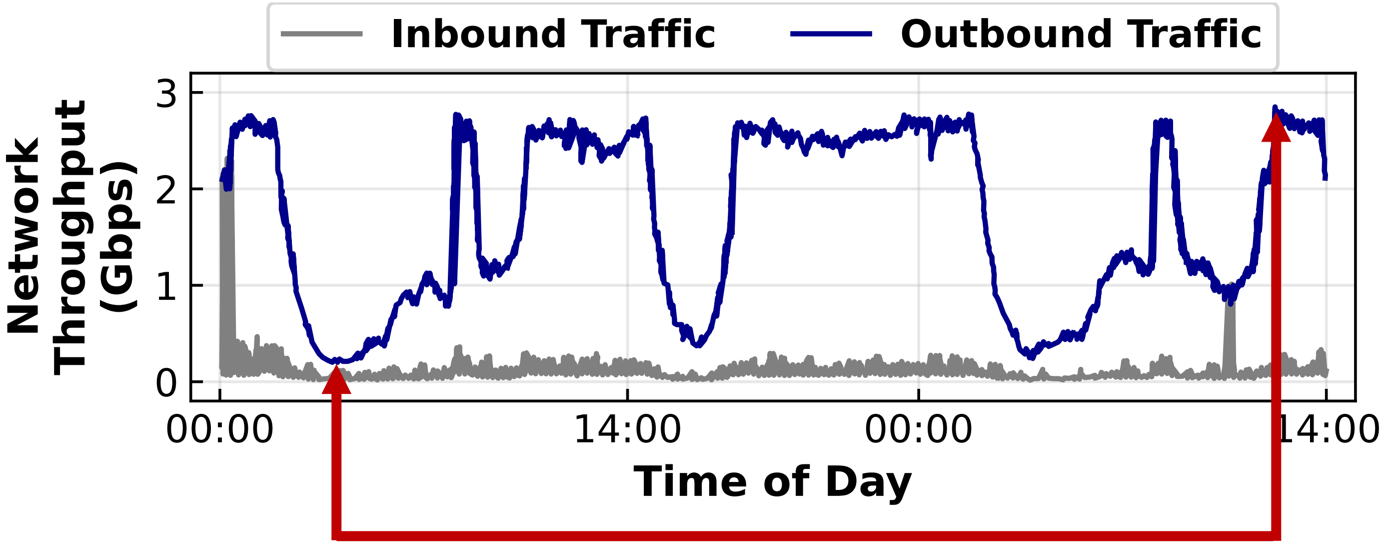


The Conceptual Architecture of an SoC Cluster

- Computing units: Every 5 mobile SoCs are integrated into one printable circuit board (PCB). (60 SoCs in total)
- Networking: One backplane and one Ethernet Switch Board (20 Gbps); 2-layer Ethernet networking and power supply
- Server management: One baseboard management controller (BMC); SoC/PCB controler, thermal manager, etc.
- Cooling devices: 8 fans

Trace Anslysis of Mobile Cloud Gaming

- Real-world traces: Network traffic of an in-the-wild SoC Cluster over 38 hours, only serving mobile cloud gaming services



Up to 25x Outbound Traffic Gap

High hardware usage variation drives us to explore **whether SoC Clusters can efficiently support other workloads.**

Micro-benchmarks on CPU

- Micro-benchmarks: Geekbench 5
- Hardware:
  - One traditional edge server with Intel Xeon 5218R CPU (40 cores)
  - AWS Graviton 2/3 cloud instances with ARM CPUs (m6/7g.metal, 64 cores)
  - An SoC Cluster (60 * 8 cores, Qualcomm Snapdragon 865 SoC)

Micro Benchmarks	Per-core Performance				Whole Server Performance			
	Ours	Trad.	G2	G3	Ours	Trad.	G2	G3
CPU Score	911	840	762	1,121	194,100	15,450	36,091	51,379
Integer Score	842	800	735	1,039	184,500	16,224	36,653	50,695
Floating Score	948	886	790	1,214	191,820	15,793	35,813	49,885
Text Compress	4.4	4.1	4.2	4.9	906	135	195	206
SQLite Query	257	249	208	279	59,958	9,240	12,200	16,200
PDF Render	52	41	37	66	12,552	710	2,140	3,960

SoC Cluster aligns closely with Trad. Intel CPU server, outperforming AWS Graviton 2 but not matching the performance of the AWS Graviton 3 instance.

Micro-benchmarks on CPU

- Micro-benchmarks: Geekbench 5
- Hardware:
  - One traditional edge server with Intel Xeon 5218R CPU (40 cores)
  - AWS Graviton 2/3 cloud instances with ARM CPUs (m6/7g.metal, 64 cores)
  - An SoC Cluster (60 * 8 cores, Qualcomm Snapdragon 865 SoC)

Micro Benchmarks	Per-core Performance				Whole Server Performance			
	Ours	Trad.	G2	G3	Ours	Trad.	G2	G3
CPU Score	911	840	762	1,121	194,100	15,450	36,091	51,379
Integer Score	842	800	735	1,039	184,500	16,224	36,653	50,695
Floating Score	948	886	790	1,214	191,820	15,793	35,813	49,885
Text Compress	4.4	4.1	4.2	4.9	906	135	195	206
SQLite Query	257	249	208	279	59,958	9,240	12,200	16,200
PDF Render	52	41	37	66	12,552	710	2,140	3,960

The large number of SoC CPU cores delivers superior performance compared to other CPU servers.

SoC Cluster Benchmark

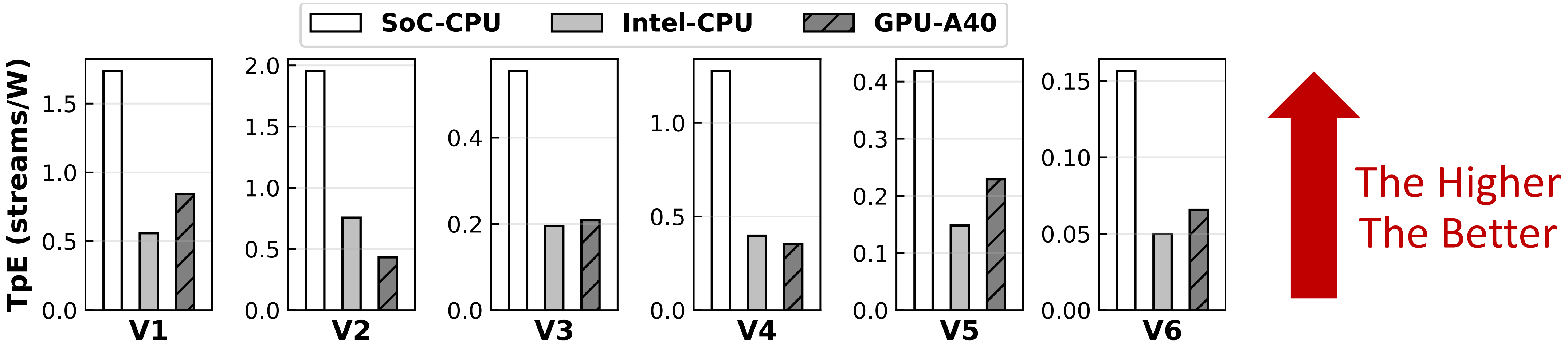
- Video transcoding: Live streaming transcoding & archive transcoding
  - Software: FFmpeg & LiTr[1]
  - Dataset: 6 videos picked from vbench[2] with diverse video complexities
  - Metrics: Throughput, energy efficiency, video bitrate, video quality
- Deep learning (DL) serving
  - Software: TVM on Intel CPU; TensorRT on NVIDIA GPU; TFLite on SoC Clusters
  - Models: ResNet-50 (FP32/INT8), ResNet-152 (FP32/INT8), YOLOv5x (FP32), BERT (FP32)
  - Metrics: Latency, throughput, energy efficiency
- Alternative Hardware
  - One physical edge server: Intel Xeon 5218R Gold Processor (40 cores)
  - Datacenter-level GPUs: NVIDIA A40 & NVIDIA A100

[1] https://github.com/linkedin/LiTr

[2] [ASPLOS'18] Andrea Lottarini et al. vbench: Benchmarking Video Transcoding in the Cloud

Video Transcoding

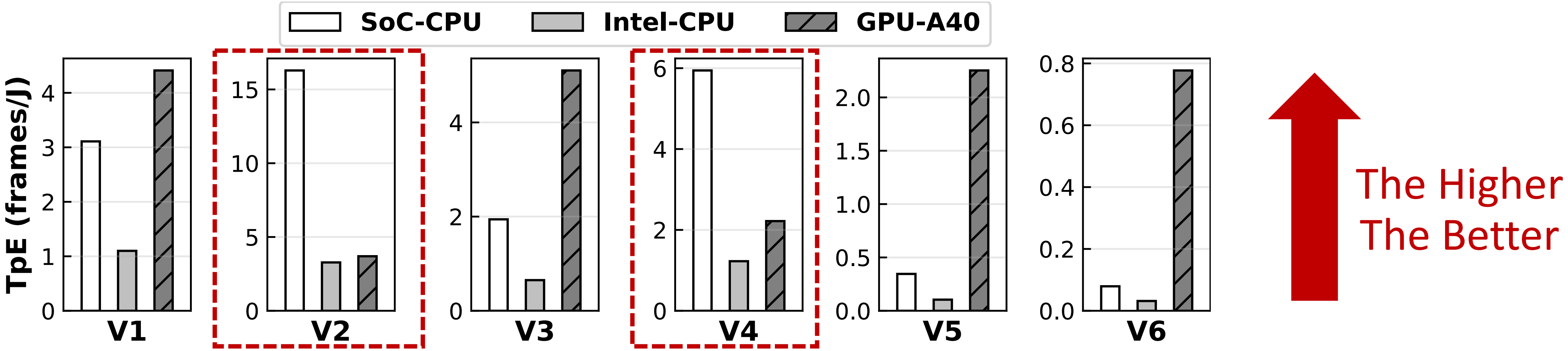
- **Question #1:** How much energy efficiency can be gained by using SoC Clusters for video transcoding?
- Task: Live streaming transcoding
- Energy efficiency: The number of streams a single watt can support



SoC CPUs are up to **3.2x** more energy-efficient than the **Intel CPU**, and up to **4.5x** more energy efficient than the **NVIDIA A40 GPU**.

Video Transcoding

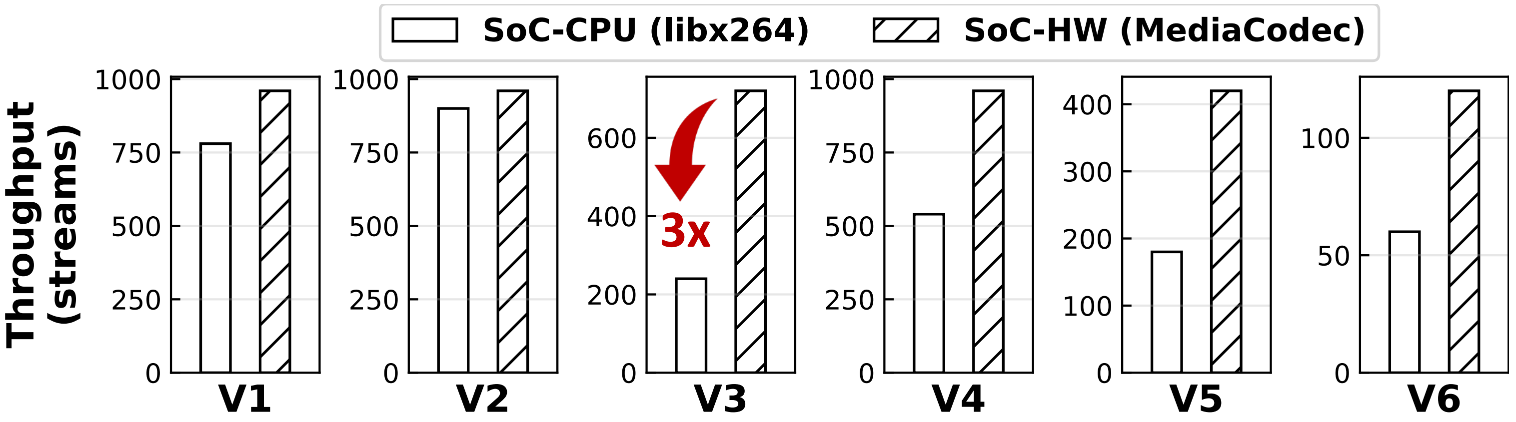
- Task: Archive transcoding (more computation required than live streaming transcoding)
- Energy efficiency: The number of frames a single Joule can process



- **SoC CPUs** still achieve higher energy efficiency than the **Intel CPU.**
- **SoC CPUs** only outperform the NVIDIA A40 GPU in **simple (low-complexity) videos** (i.e. V2 and V4), but fails in more complex ones.

Video Transcoding

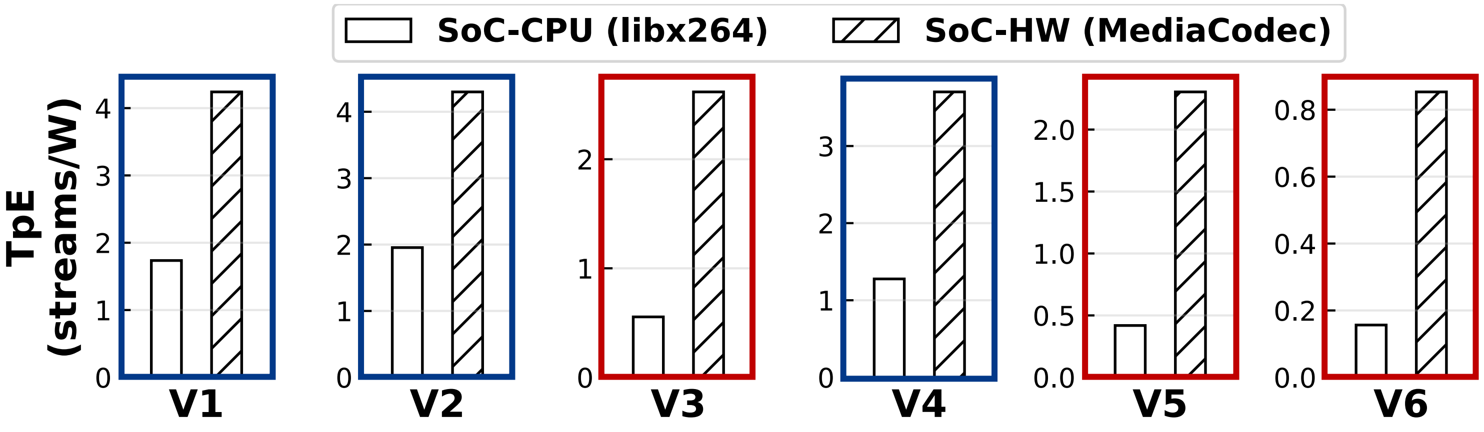
- **Question #2:** To what extent do SoC codecs outperform SoC CPUs?
- Task: Live streaming transcoding
- Metrics
  - Throughput: The number of streams a whole SoC Cluster can support
  - Energy efficiency: The number of streams a single watt can support



- Throughput: **1.07x – 3x** improvement.

Video Transcoding

- **Question #2:** To what extent do SoC codecs outperform SoC CPUs?
- Task: Live streaming transcoding
- Metrics
  - Throughput: The number of streams a whole SoC Cluster can support
  - Energy efficiency: The number of streams a single watt can support

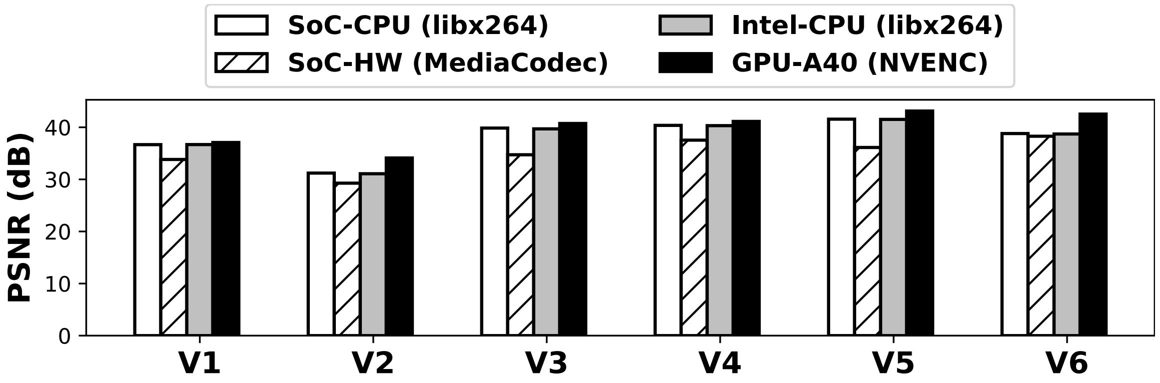


- Throughput: **1.07x – 3x** improvement.
- Energy efficiency:
  - Simple videos (**V1/2/4**): A geometric mean of **2.5x** improvement.
  - Complex videos (**V3/5/6**): **4.7x – 5.5x** improvements.

The huge potential of running live streaming transcoding on SoC Clusters!

Video Transcoding

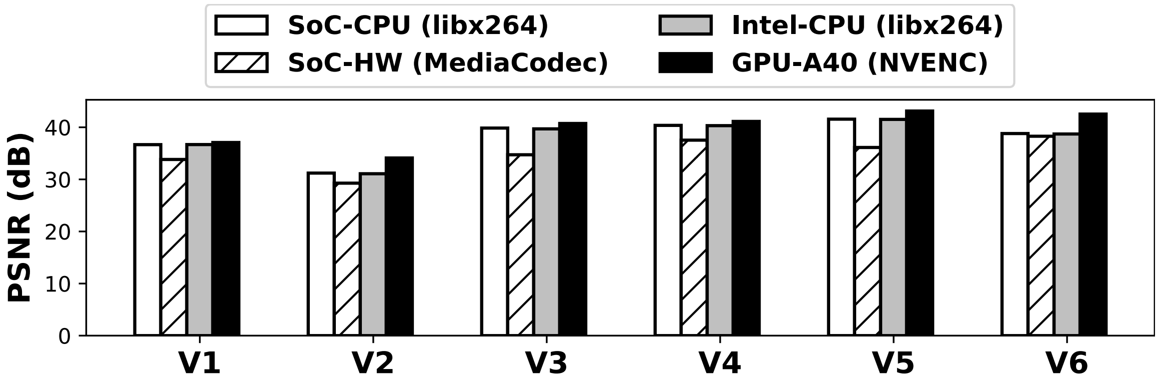
- **Question #3:** Can SoC hardware codec deliver satisfactable QoE in live video transcoding tasks?
- Metric: Video quality and video bitrate

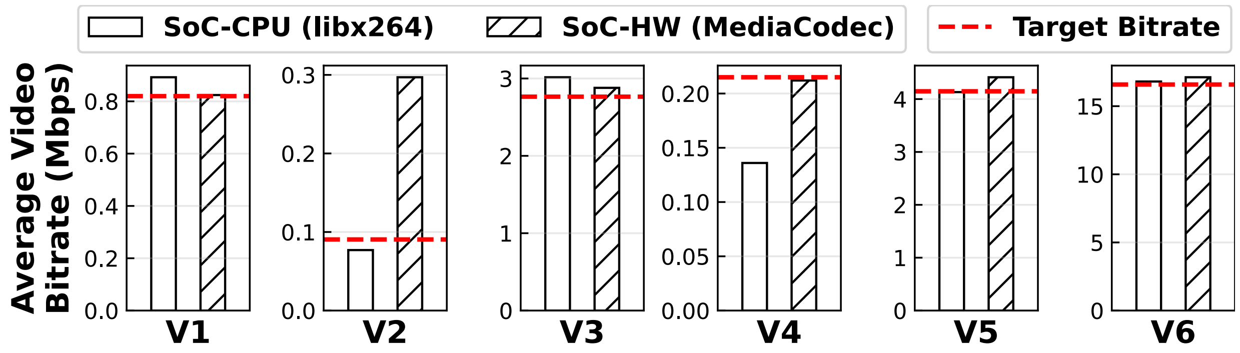


- Videos transcoded by SoC hardware codecs show up to 15% lower PSNR values.

Video Transcoding

- **Question #3:** Can SoC hardware codec deliver satisfactable QoE in live video transcoding tasks?
- Metric: Video quality and video bitrate

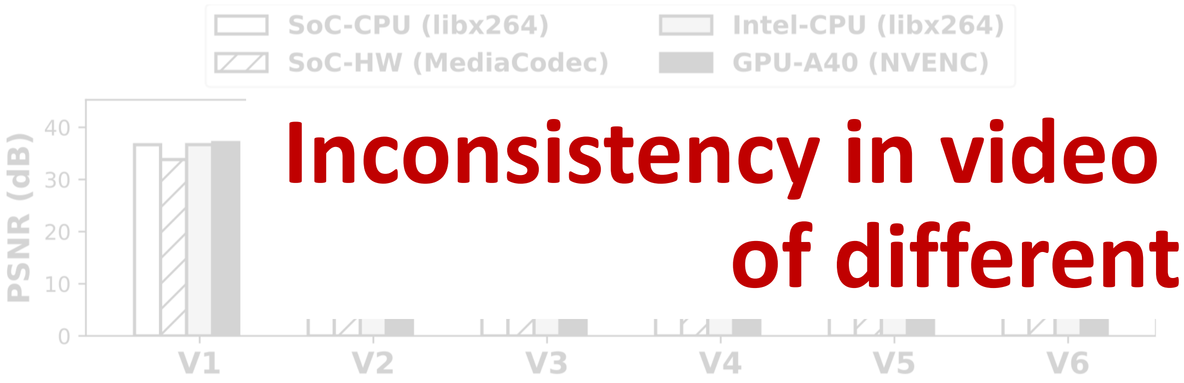




- Videos transcoded by SoC hardware codecs exhibit up to 15% lower PSNR values.
- SoC hardware codecs struggle to meet a ralatively low bitrate cap (V2).

Video Transcoding

- **Question #3:** Can SoC hardware codec deliver satisfactable QoE in live video transcoding tasks?
- Metric: Video quality and video bitrate

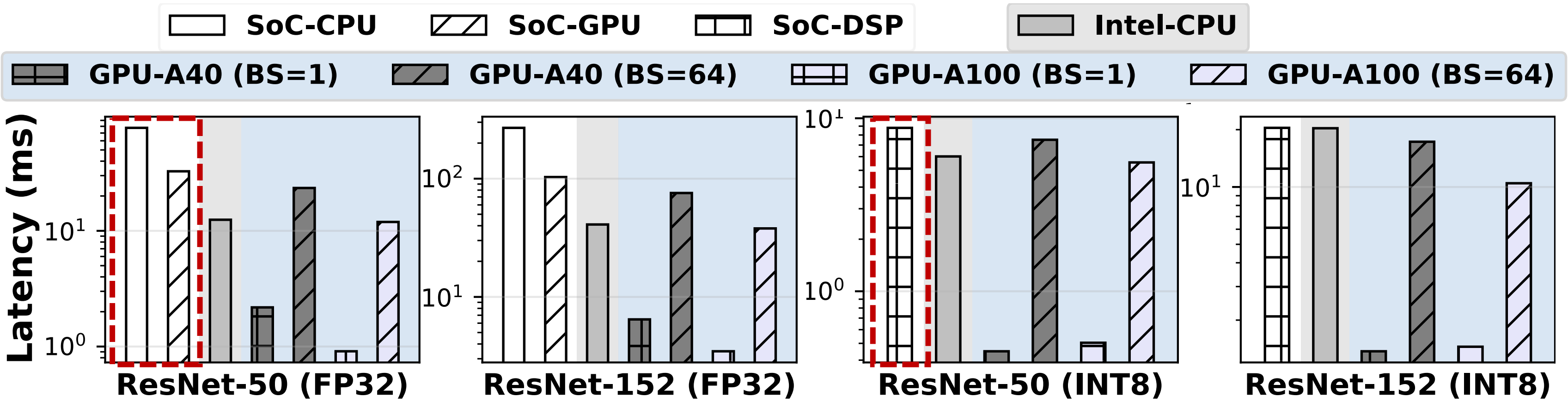


Inconsistency in video quality and video bitrate of different video codecs

Edge service oprators should judiciously select the appropriate hardware to meet app QoE, if using SoC Clusters in live streaming transcoding tasks.

Deep Learning Serving

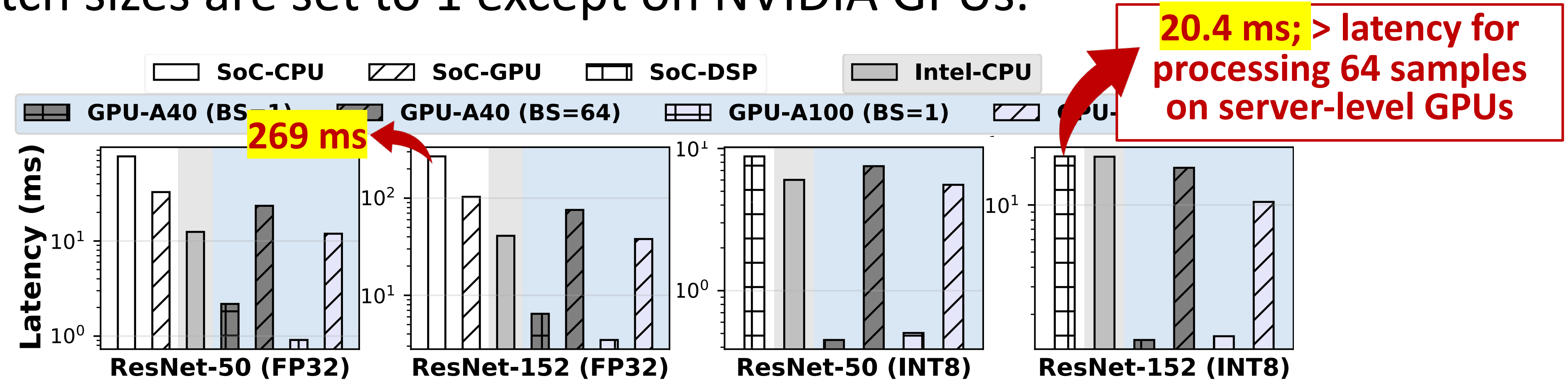
- **Question #1:** Can SoC Clusters support DL serving workloads with low latency?
- All batch sizes are set to 1 except on NVIDIA GPUs.



- SoC DSPs could deliver an adequate latency in medium-sized DNN, e.g., 8.8ms on quantized ResNet-50.

Deep Learning Serving

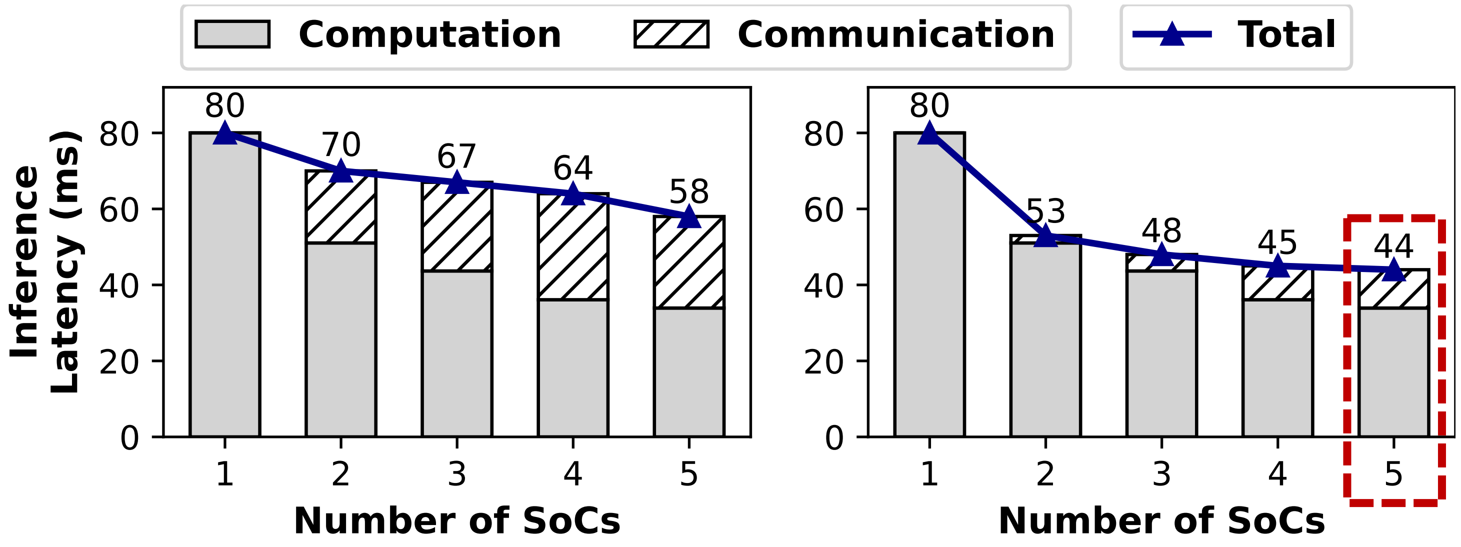
- **Question #1:** Can SoC Clusters support DL serving workloads with low latency?
- All batch sizes are set to 1 except on NVIDIA GPUs.



- SoC DSPs could deliver an adequate latency in medium-sized DNN, e.g., 8.8 ms on quantized ResNet-50.
- SoC Clusters challenge to handle large models with individual SoCs.

Deep Learning Serving

- **Question #2:** Can involving more SoCs for SoC-collaborative inference deliver low latency on large models?
- Model: ResNet-50 (FP32)
- Approach: (Left) Tensor parallelism proposed in CoEdge[1]; (Right) Tensor parallelism with computation/communication pipelining

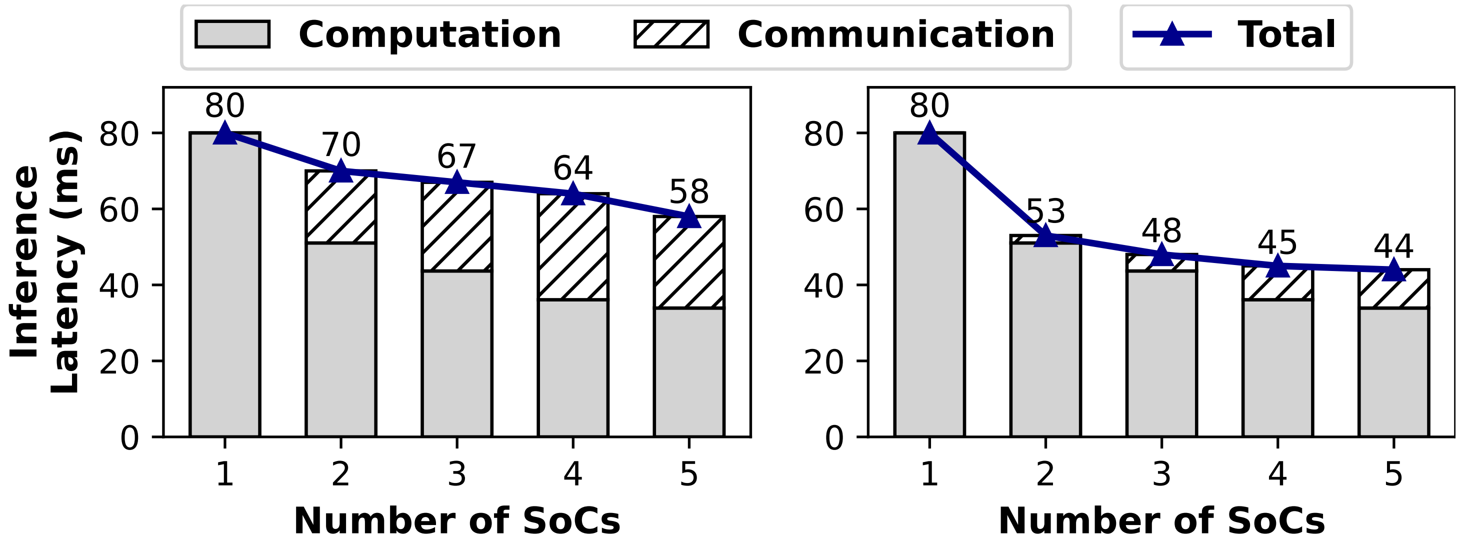


- Involving more SoCs does not proportionally reduce inference latencies.
- Even with the optimized software, network communication time still accounts for 23% (5 SoCs).

[1] Zeng et al., CoEdge: Cooperative DNN Inference With Adaptive Workload Partitioning Over Heterogeneous Edge Devices

Deep Learning Serving

- **Question #2:** Can involving more SoCs for SoC-collaborative inference deliver low latency on large models?
- Model: ResNet-50 (FP32)
- Approach: (Left) Tensor parallelism proposed in CoEdge[1]; (Right) Tensor parallelism with computation/communication pipelining



Software enhancements (e.g., more fine-grained tensor partitioning) and hardware enhancements (e.g., improving network bandwidth) should be utilized jointly.

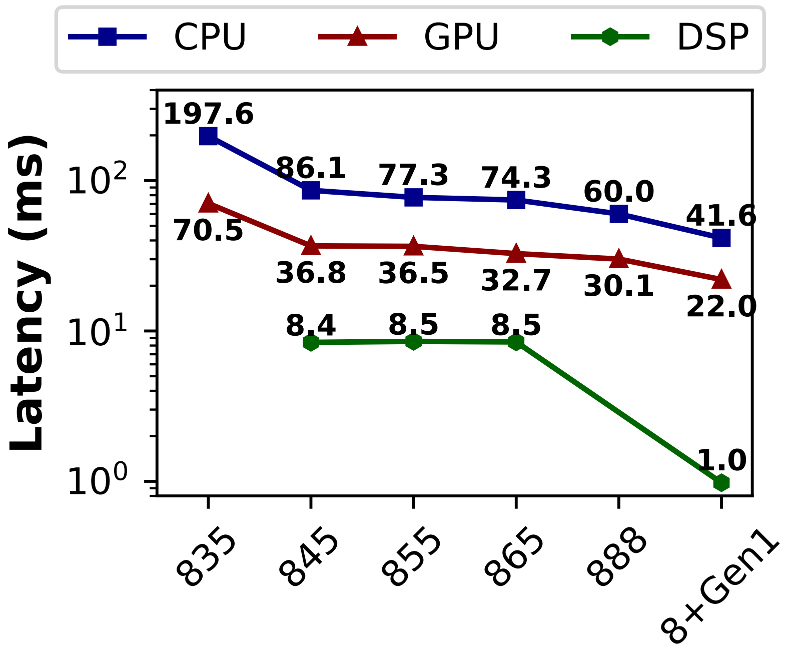
[1] Zeng et al., CoEdge: Cooperative DNN Inference With Adaptive Workload Partitioning Over Heterogeneous Edge Devices

SoC Longitudinal Study

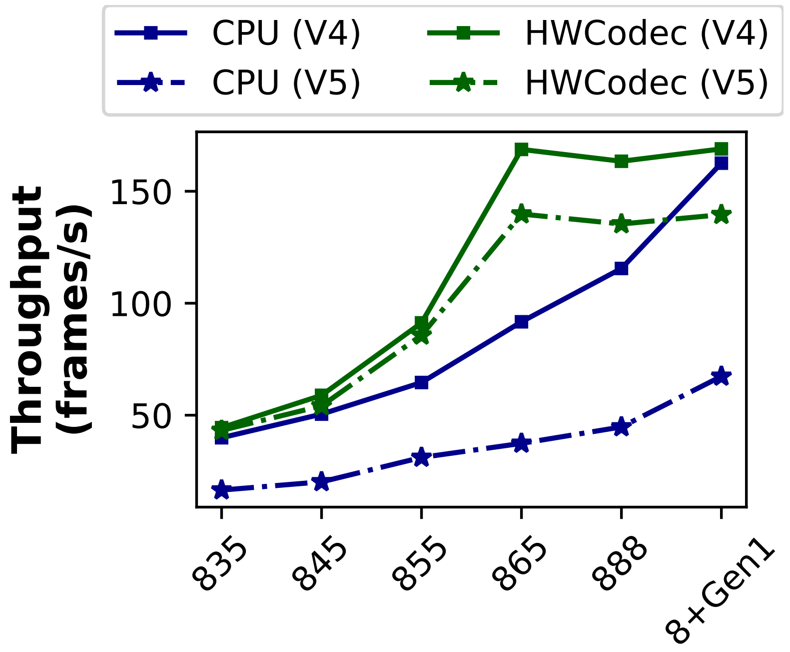
- Six Qualcomm Snapdragon 8-series SoC models (2017 – 2022)
- Two workloads: DL serving and live streaming transcoding
- Metric: Latency and throughput

Devices	SoC	RAM	OS	Release Date
Xiaomi 12 S	QS 8+Gen1	12 GB	Android 12	May 2022
Xiaomi 11 Pro	QS 888	8 GB	Android 11	Jun. 2021
Meizu 17	QS 865	8 GB	Android 10	Mar. 2020
Meizu 16T	QS 855	6 GB	Android 9	Mar. 2019
Xiaomi 8	QS 845	6 GB	Android 8.1	Feb. 2018
Xiaomi 6	QS 835	6 GB	Android 7.1.1	Mar. 2017

SoC Longitudinal Study



DL Serving



Live Streaming Transcoding

- Tremendous performance improvements in the past six years.
- Mobile SoCs are promising candidates for more complex server-side workloads.
- Leverage the co-processors to fully unleash their performance.

Conclusion

- Energy efficiency is critical to edge platforms.
- An extreme design towards energy efficiency: SoC Cluster
  - Massive low-power mobile processors
  - Every SoC is inherently heterogeneous (with GPU/NPU/video codec)
  - Commercial success in mobile cloud gaming services
- A set of experiments to demonstrate the pros/cons of SoC Cluster over traditional servers.
  - More experiments and results in our paper!
- Show potential directions for software- and hardware-level optimizations in the future.

- ❑ Benchmark suite: https://github.com/SoC-Cluster/SoC-Cluster-artifacts
- ❑ Online access to cloud phone services powered by SoC Clusters: https://www.alibabacloud.com/help/en/ecp/what-is-ecp
- ❑ Contact: li.zhang@bupt.edu.cn Website: https://lizhang20.github.io

More is Different: Prototyping and Analyzing a New Form of Edge Server with Massive Mobile SoCs

Li Zhang[1], Zhe Fu[2], Boqing Shi[1], Xiang Li[1], Rujin Lai[3], Chenyang Yang[3]
Ao Zhou[1], Xiao Ma[1], Shangguang Wang[1], Mengwei Xu[1]
[1]*Beijing University of Posts and Telecommunications*
[2]*Tsinghua University*, [3]*vclusters*

Thank you!

Happy to take questions about this new type of edge server!