A Satellite-Born Server Design with Massive Tiny Chips Towards In-Space Computing

Mengwei Xu, Li Zhang, Hongyu Li, Ruolin Xing, Qibo Sun

Beijing University of Posts and Telecommunications

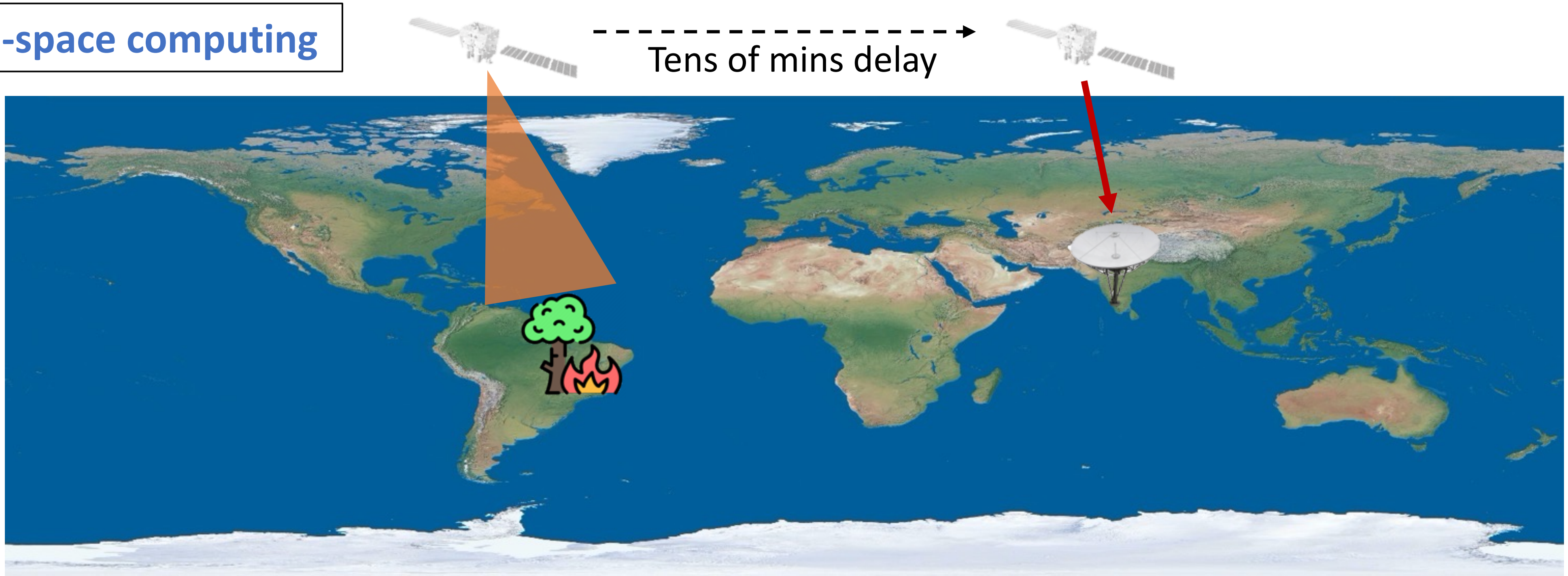
Background: In-Space Computing

- Process space-native data such as earth imagery
  - Abundant: satellite-ground link can only support a tiny portion of total data (TBs per day) downloaded for post-processing
  - Realtime: deliver post-processing messages to the ground through GEO like Beidou

Background: In-Space Computing

- Process space-native data such as earth imagery
  - Abundant: satellite-ground link can only support a tiny portion of total data (TBs per day) downloaded for post-processing
  - Realtime: deliver post-processing messages to the ground through GEO like Beidou

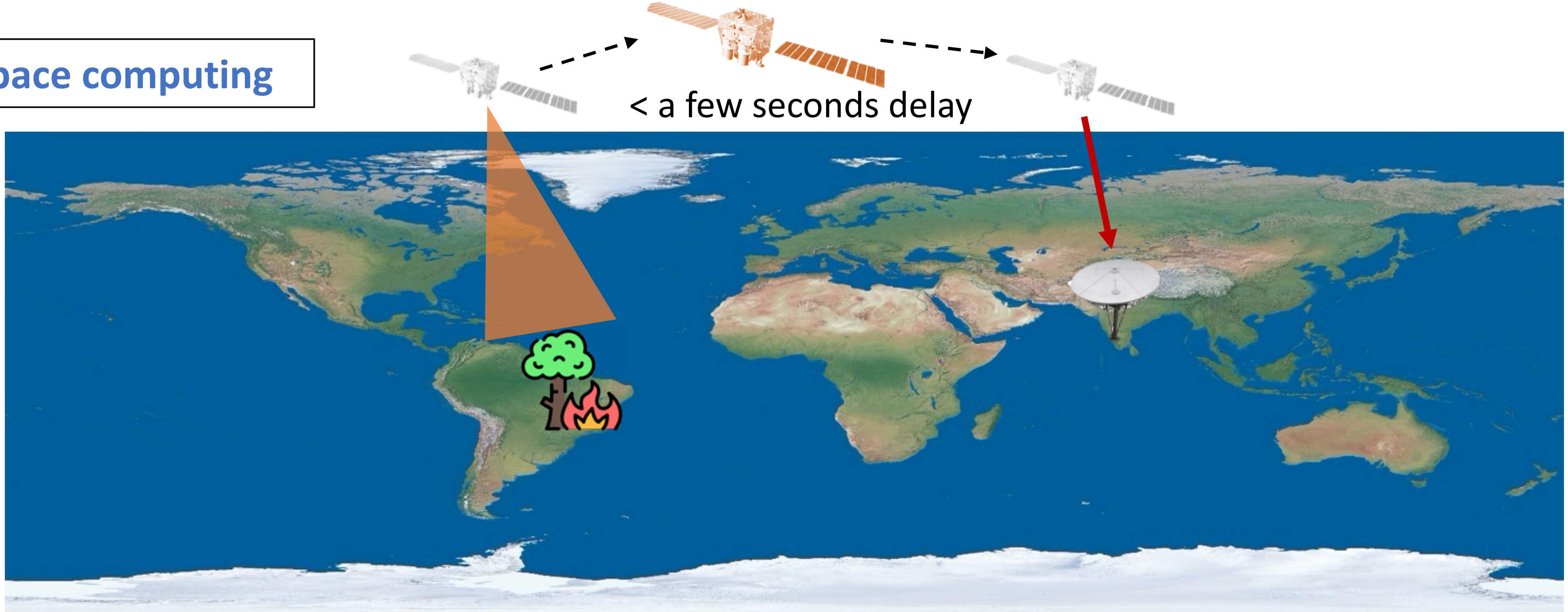
Without in-space computing



Background: In-Space Computing

- Process space-native data such as earth imagery
  - Abundant: satellite-ground link can only support a tiny portion of total data (TBs per day) downloaded for post-processing
  - Realtime: deliver post-processing messages to the ground through GEO like Beidou

With in-space computing

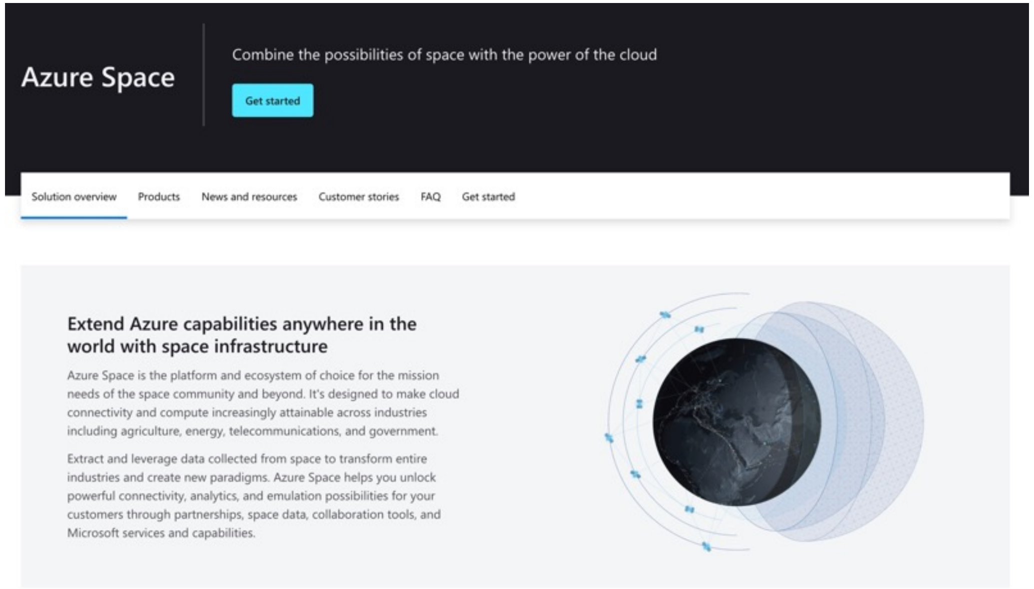


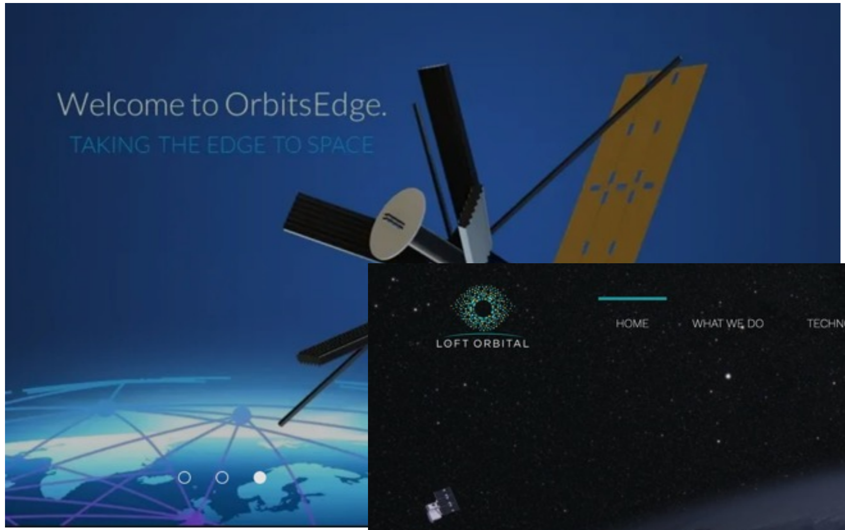
Background: In-Space Computing

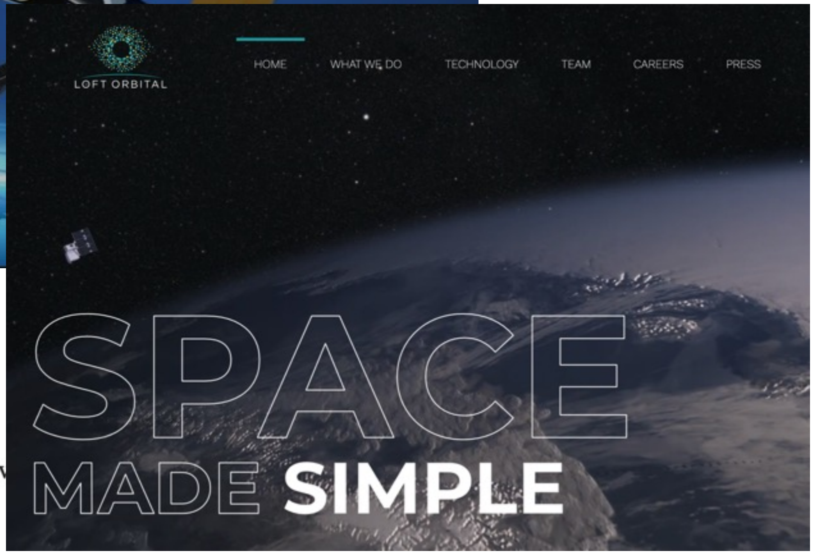
- Process space-native data such as earth imagery
  - Abundant: satellite-ground link can only support a tiny portion of total data (TBs per day) downloaded for post-processing
  - Realtime: deliver post-processing messages to the ground through GEO like Beidou
- Process offloaded data/tasks from ground as an edge node
  - High availability: anywhere and anytime (Starlink); do not suffer from geological disasters such as earthquakes
  - Green energy: zero carbon as satellites operate on harvested solar energy

Background: In-Space Computing

- Cloud computing companies: "We Need Cloud Computing in Space!"
- Startups like OrbitsEdge: "We Need Edge Computing In Space!"





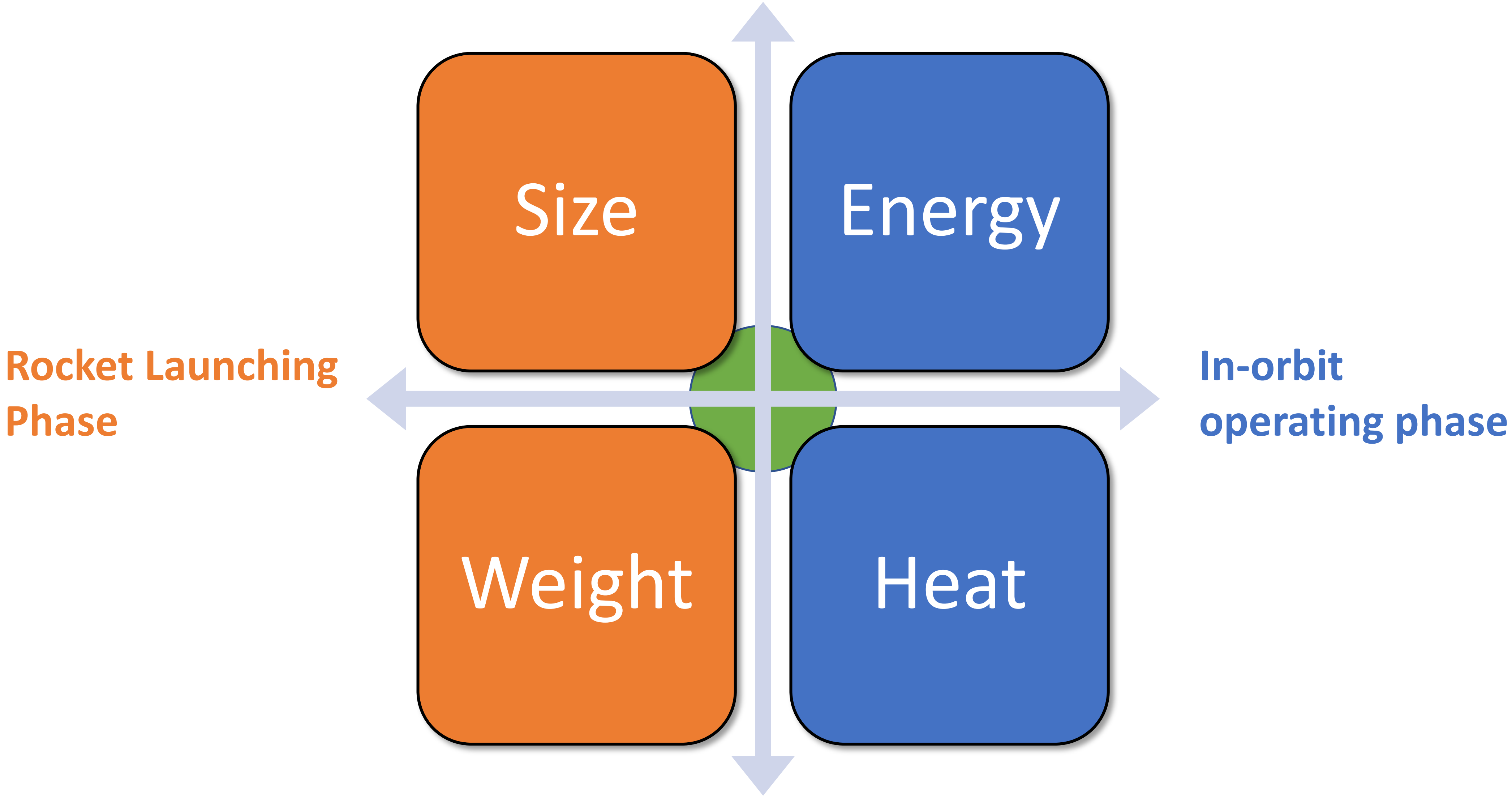


The Space Cloud: Satellite Strategies for AWS, Google and Microsoft

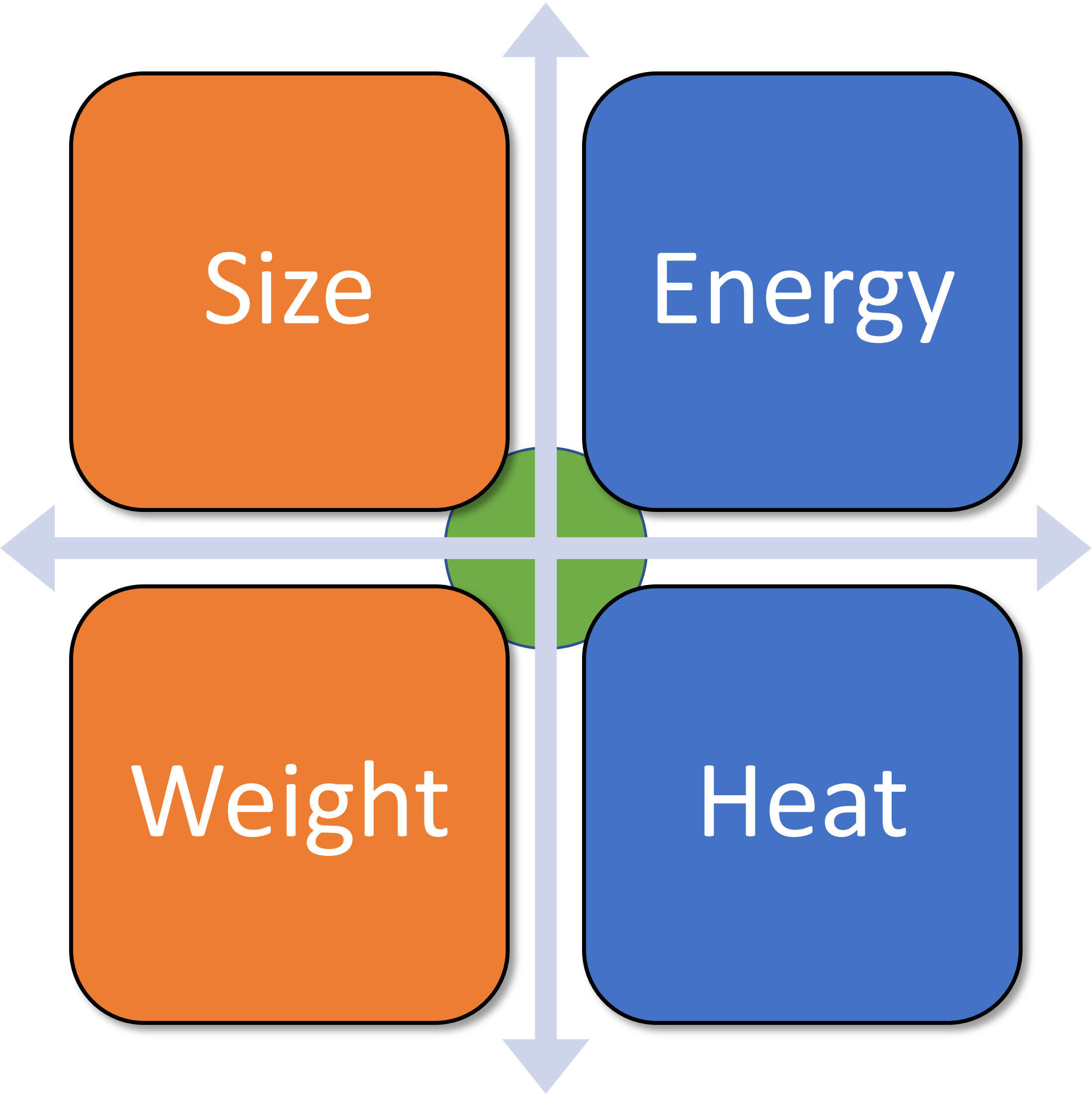
Cloud providers Amazon, Google, and Microsoft are building relationships with satellite companies. Here's a closer look at the distinctiv
strategies these cloud players are adopting as they seek to built two-way business connections between the cloud and space.

Doug Mohney

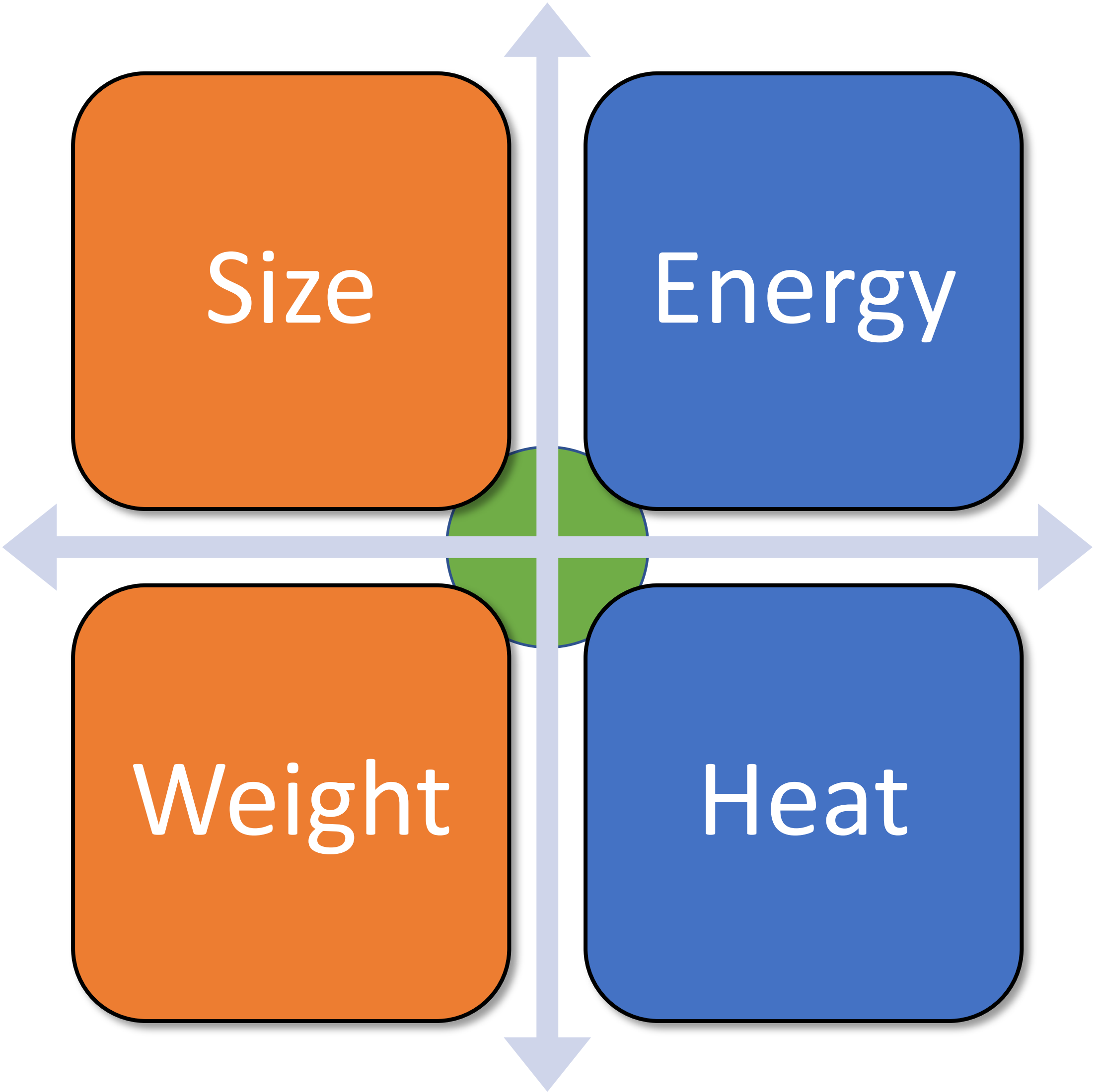
In-Space Computing Constraints



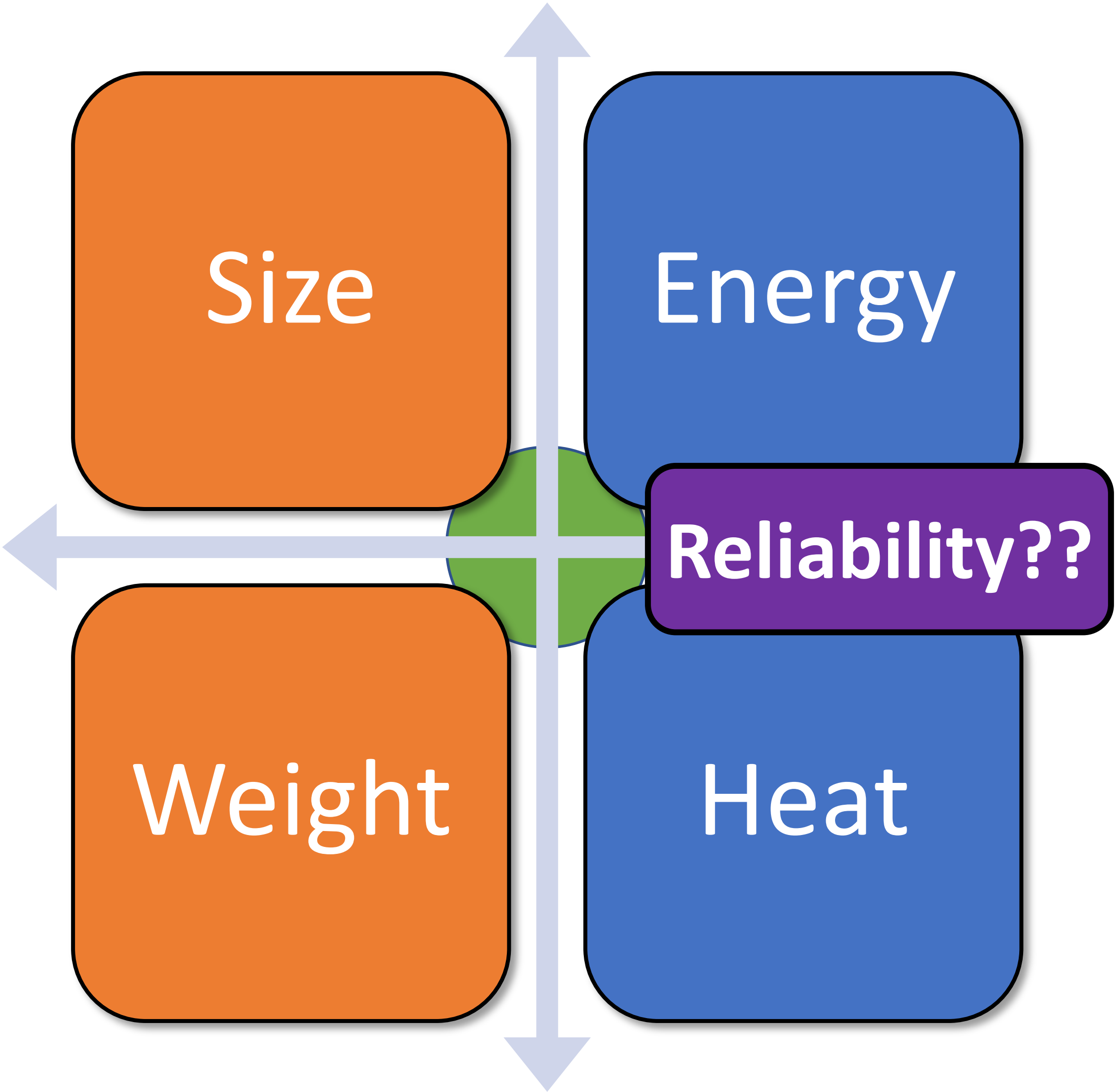
The best HW candidate?



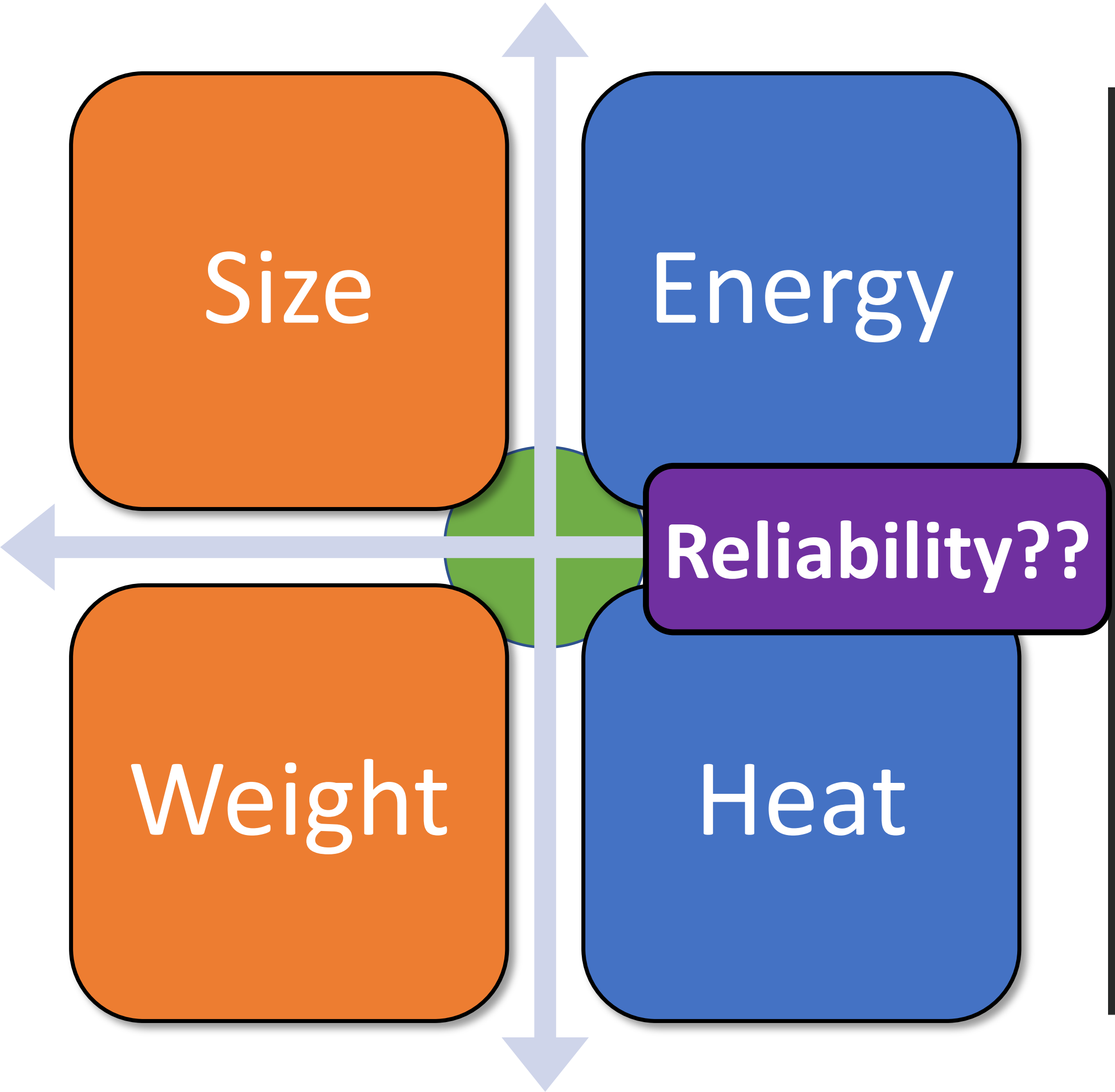
Smartphone!

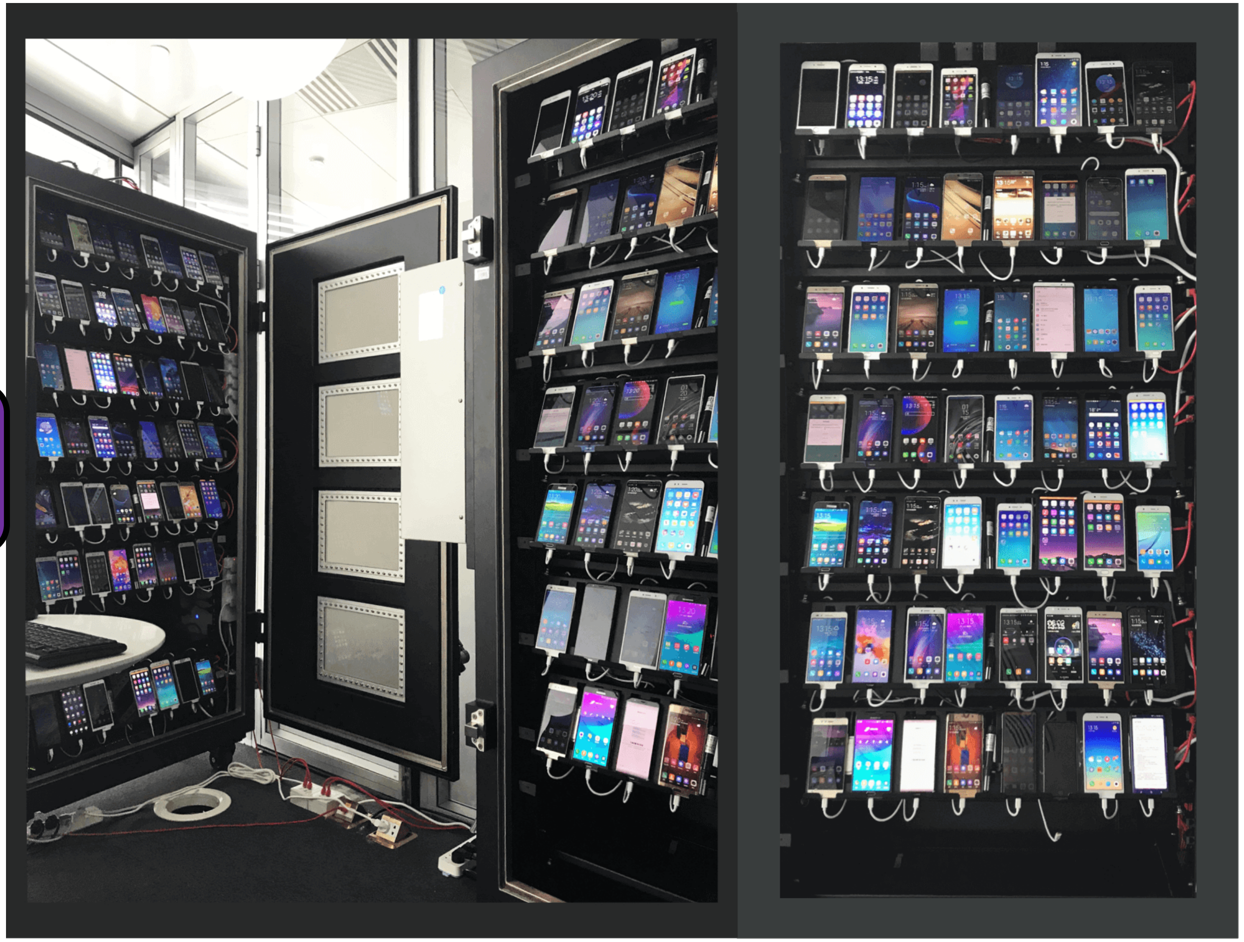


Smartphone!

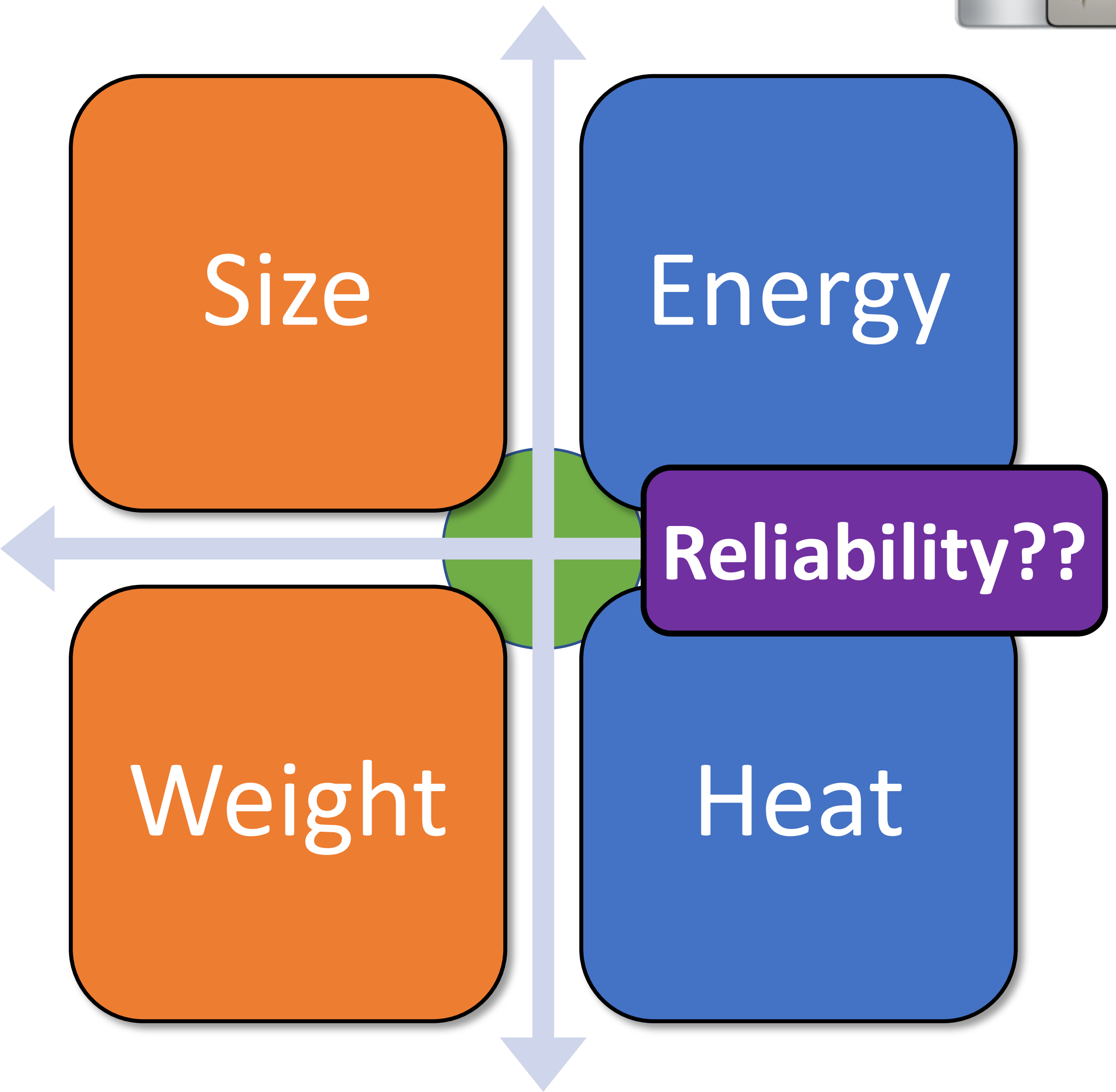


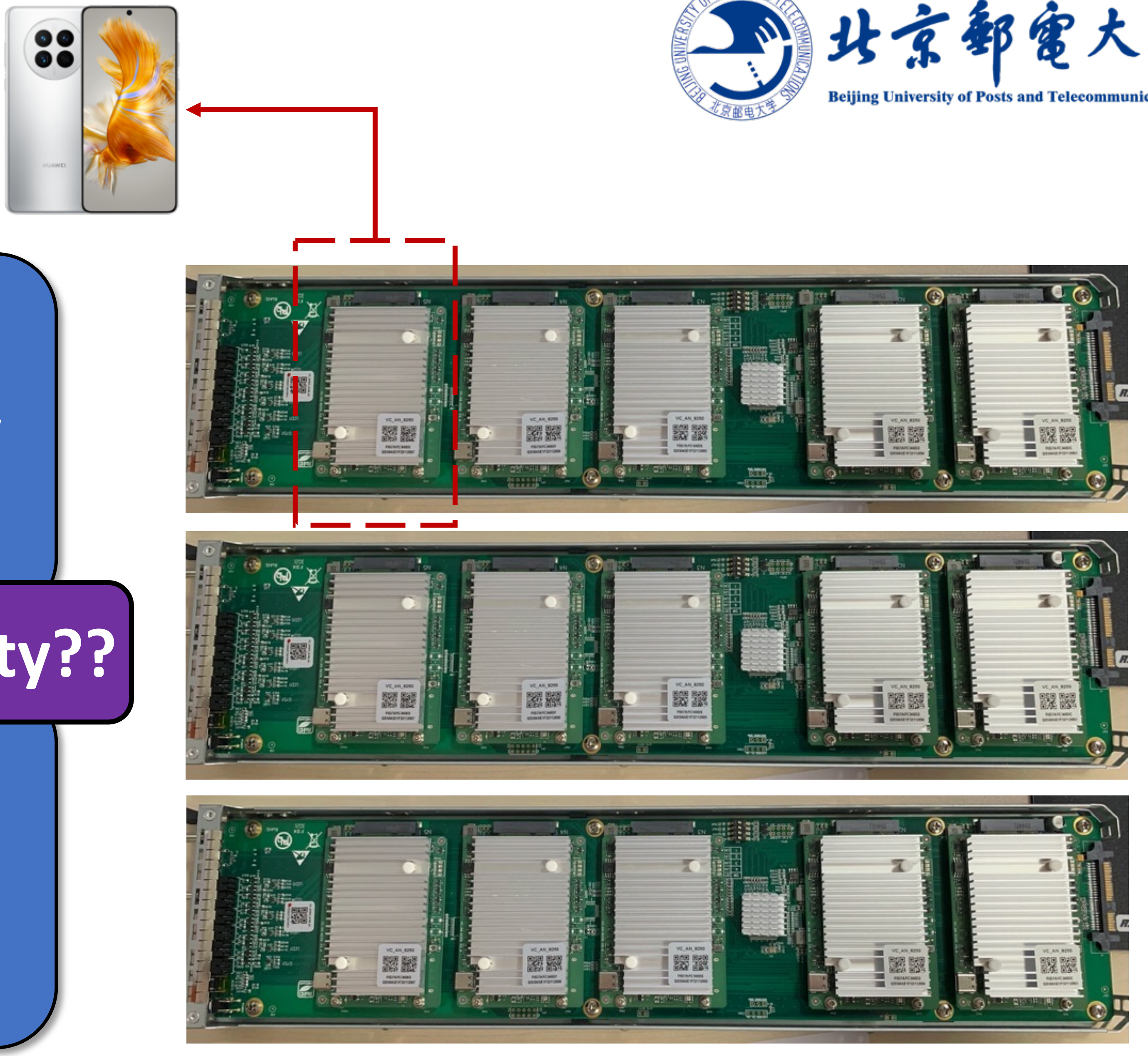
Many smartphones!



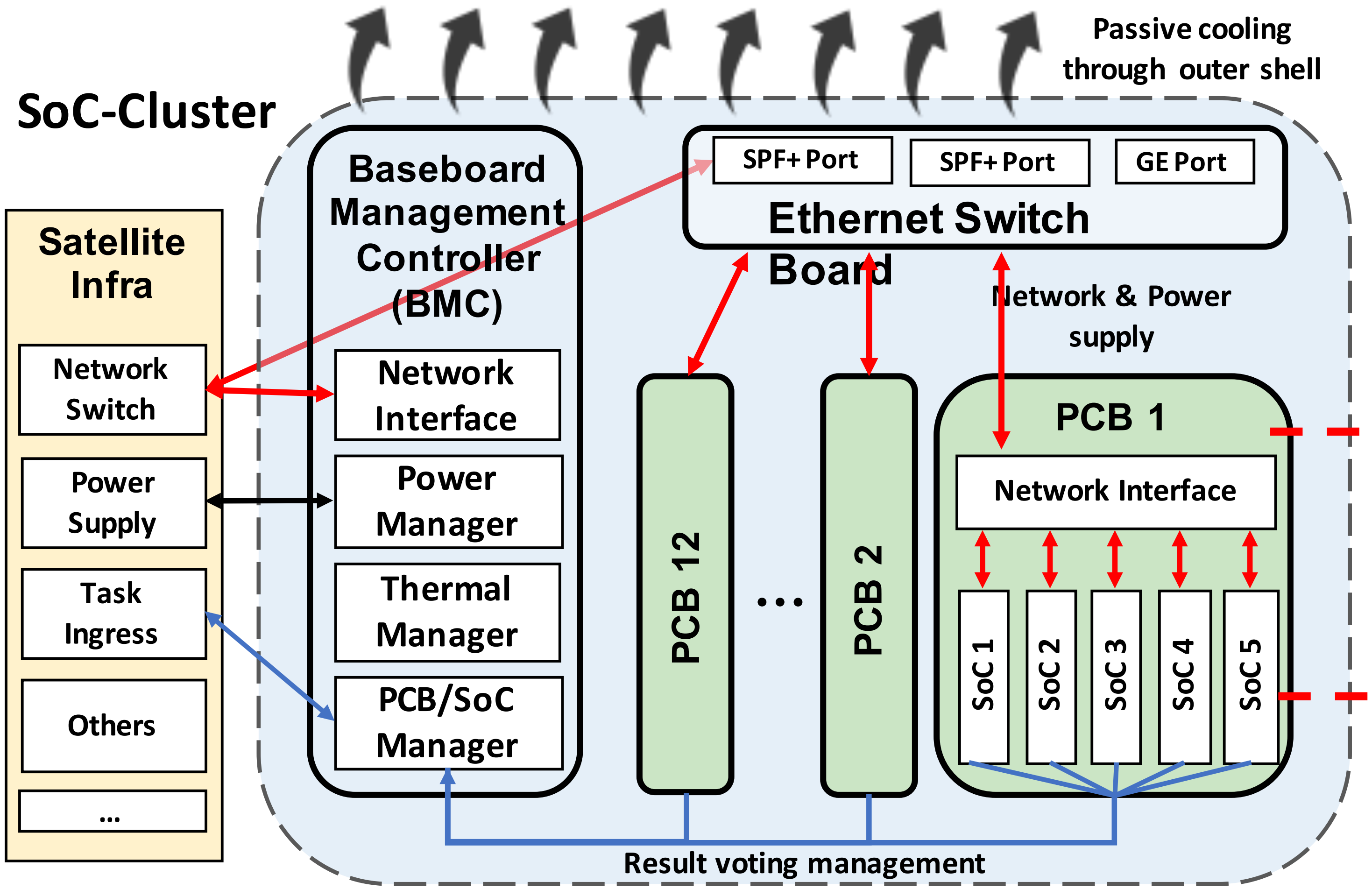


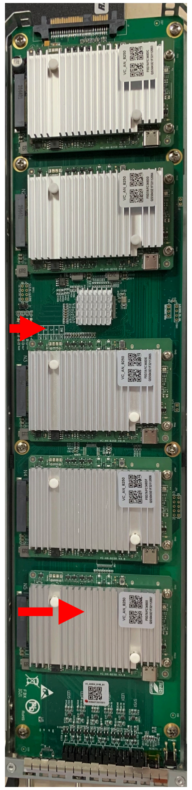
Many SoCs!



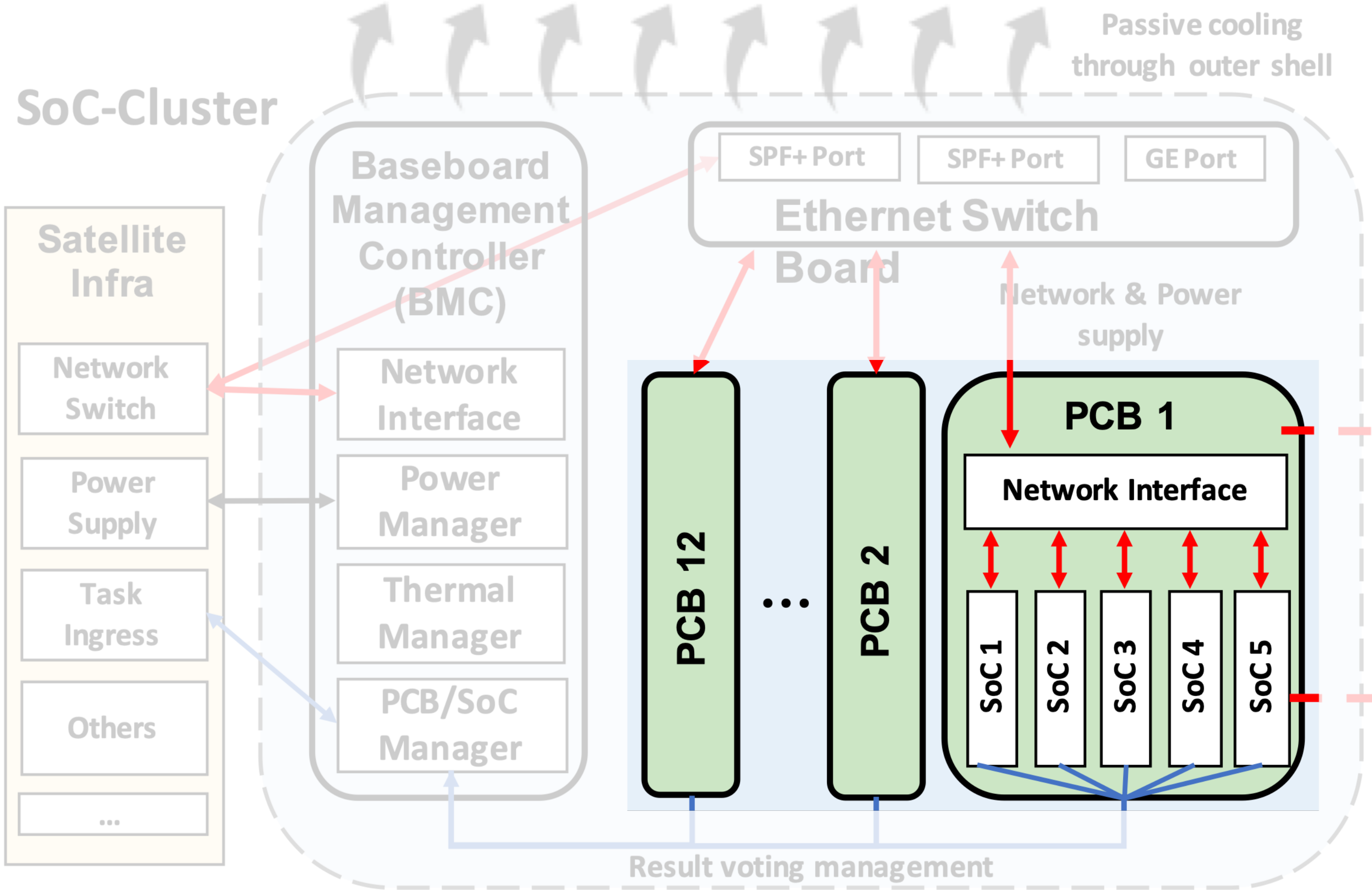


Our proposal of satellite server



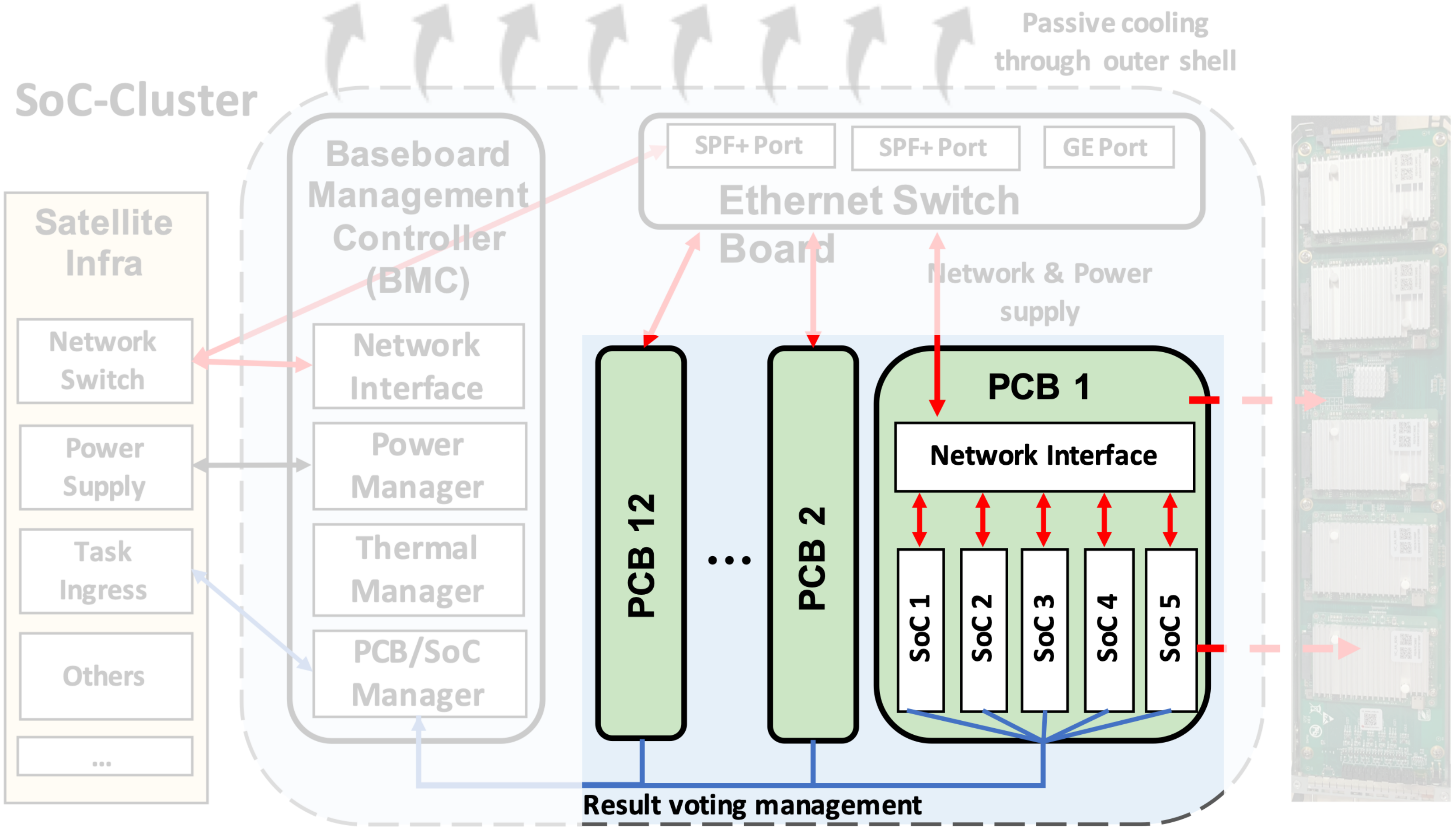


Our proposal of satellite server



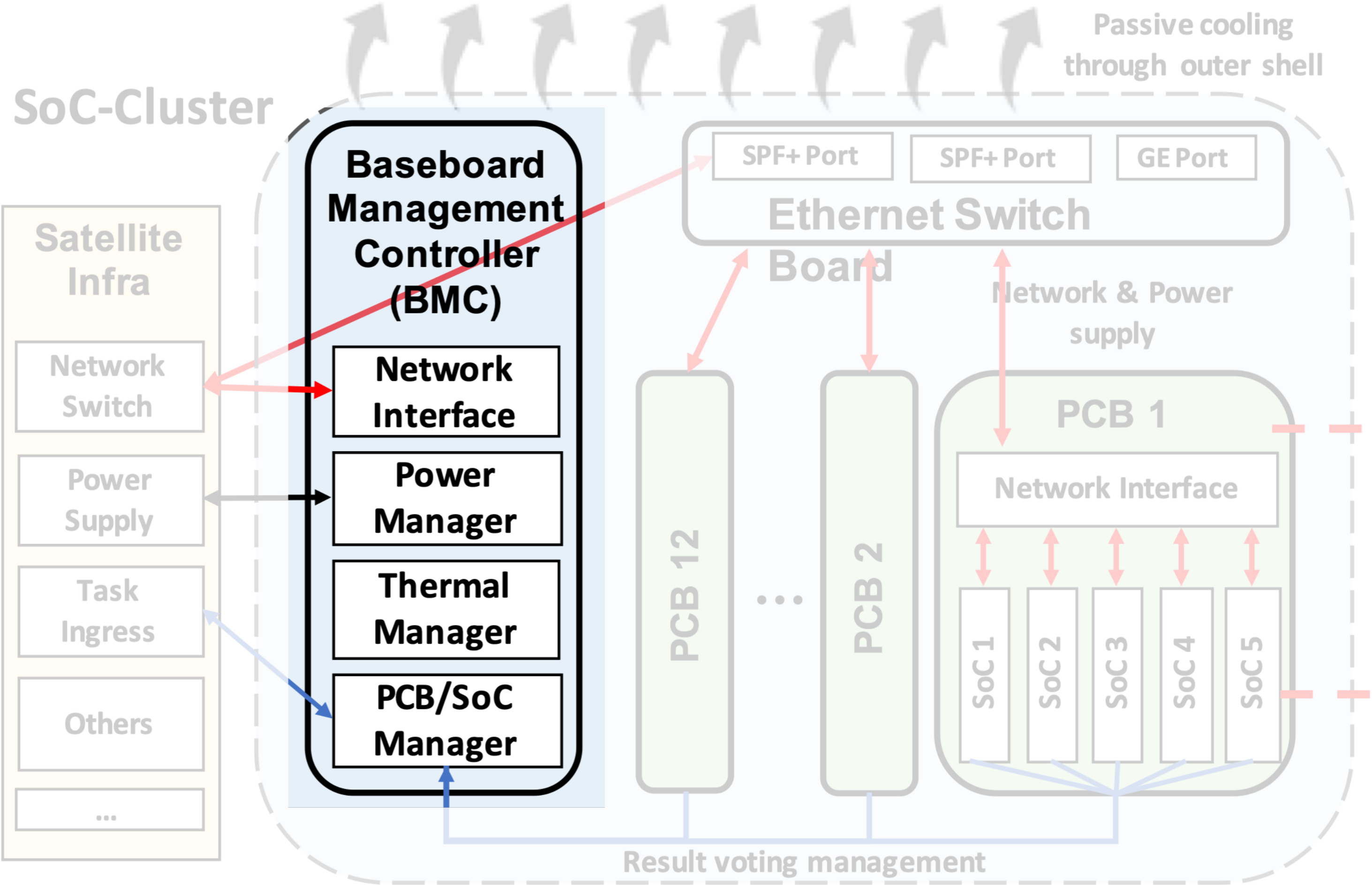
- Missive, tiny, sub-10 nm mobile SoCs
  - 60 in 2U rack

Our proposal of satellite server



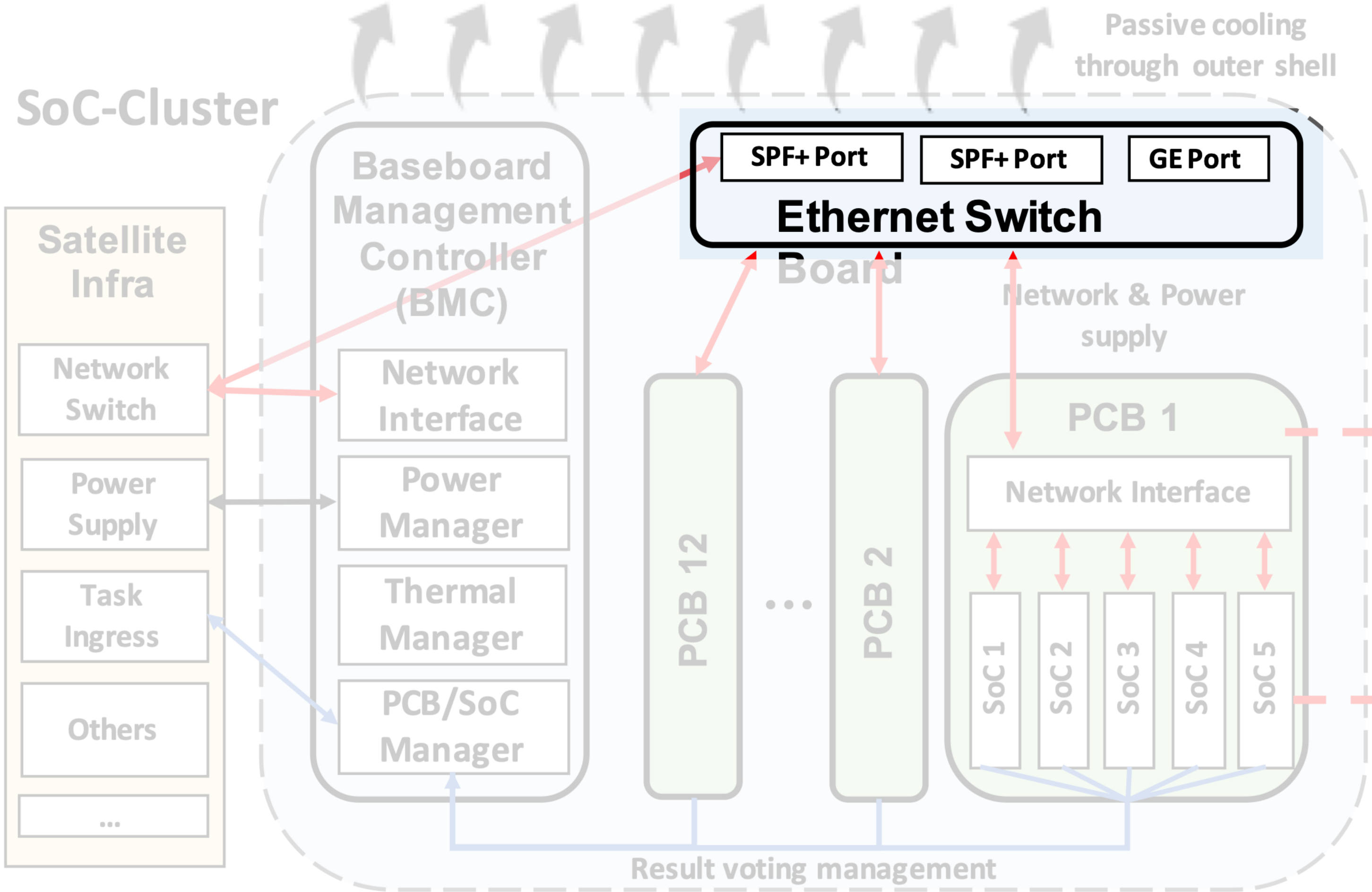
- Missive, tiny, sub-10 nm mobile SoCs
  - 60 in 2U rack
- Reliability: critical tasks run many SoCs and go through a majority voting
  - Flexible tradeoffs

Our proposal of satellite server



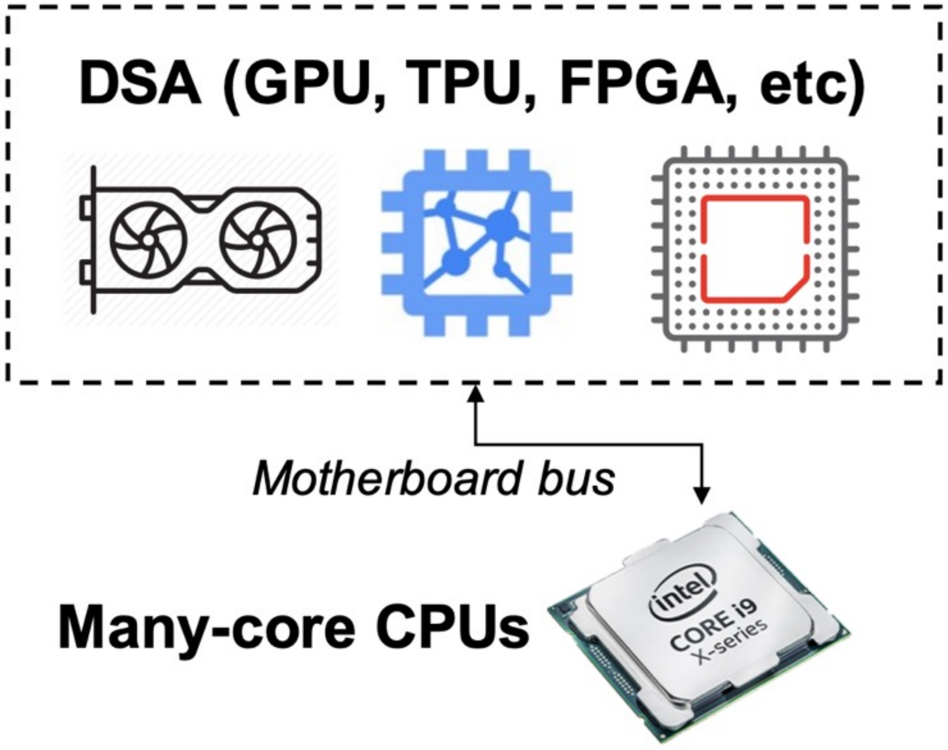
- Missive, tiny, sub-10 nm mobile SoCs
  - 60 in 2U rack
- Reliability: critical tasks run many SoCs and go through a majority voting
  - Flexible tradeoffs
- BMC: managing the whole board, scheduling tasks, hardened

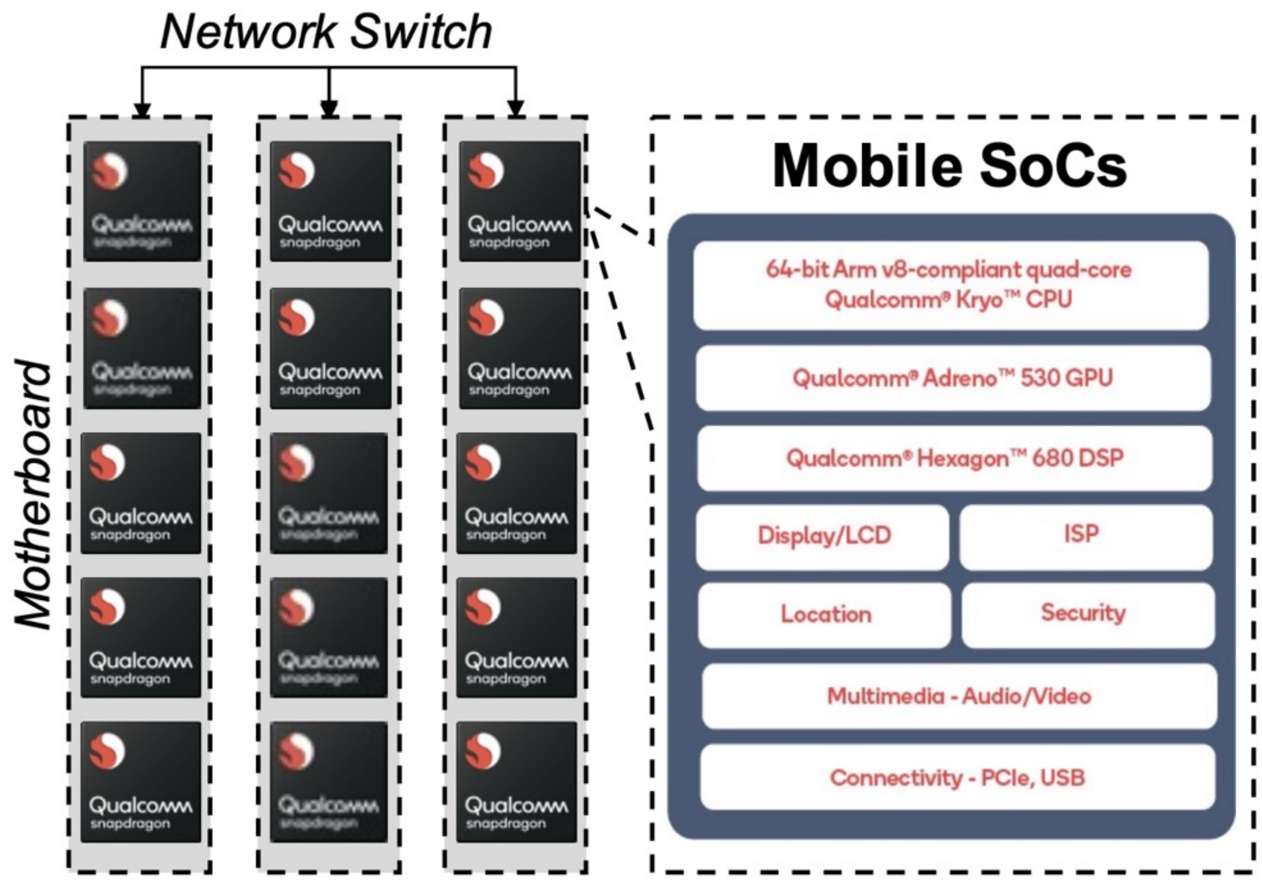
Our proposal of satellite server



- Missive, tiny, sub-10 nm mobile SoCs
  - 60 in 2U rack
- Reliability: critical tasks run many SoCs and go through a majority voting
  - Flexible tradeoffs
- BMC: managing the whole board, scheduling tasks, hardened
- Connecting SoCs and BMC through a standard ethernet switch

A high level comparison





Traditional servers

Monolithic

Our satellite server

Decentralized

Comparing satellite servers

	Throughput per Energy (TpE)			Throughput per Volume (TpV)			Throughput per Weight (TpW)		
	Power (watt)	GFLOPs per watt (FP32)	GINOPs per watt (INT8)	Volume (U)	GFLOPs per U (FP32)	GINOPs per U (INT8)	Weight (kg)	GFLOPs per kg (FP32)	GINOPs per kg (INT8)
Xeon 40-core CPU Server	276.3	0.8	0.5	1	208.3	130.4	18.8	11.1	6.9
NVIDIA A40 GPU Server	2,000.0	149.6	1,197.2	4	74,800	598,600	57.9	5,165.8	41,339.8
PowerEdge R350	95.0	0.5	0.9	1	49.3	85.4	13.6	3.6	6.3
PowerEdge R550	330.0	0.5	0.9	2	83.0	151.3	20.4	8.1	14.8
PowerEdge R750xs	370.0	0.6	1.0	2	104.6	182.5	21.9	9.5	16.6
SoC-Cluster (Kryo CPU)	672.0	1.3	0.2	2	437.4	76.5	27.0	32.4	5.7
SoC-Cluster (Adreno GPU)	387.0	193.8	X	2	37,500	X	27.0	2,777.8	X
SoC-Cluster (Hexagon DSP)	345.5	X	2,604.9	2	X	450,000	27.0	X	33,333.3

TABLE I

THEORETICAL COMPARISON BETWEEN SOC-CLUSTER AND CONVENTIONAL COTS EDGE SERVERS. "X" MEANS THAT THIS NUMERICAL OPERATION IS NOT SUPPORTED BY THE HARDWARE.

Comparing satellite servers

	Throughput per Energy (TpE)			Throughput per Volume (TpV)			Throughput per Weight (TpW)		
	Power (watt)	GFLOPs per watt (FP32)	GINOPs per watt (INT8)	Volume (U)	GFLOPs per U (FP32)	GINOPs per U (INT8)	Weight (kg)	GFLOPs per kg (FP32)	GINOPs per kg (INT8)
Xeon 40-core CPU Server	276.3	0.8	0.5	1	208.3	130.4	18.8	11.1	6.9
NVIDIA A40 GPU Server	2,000.0	149.6	1,197.2	4	74,800	598,600	57.9	5,165.8	41,339.8
PowerEdge R350	95.0	0.5	0.9	1	49.3	85.4	13.6	3.6	6.3
PowerEdge R550	330.0	0.5	0.9	2	83.0	151.3	20.4	8.1	14.8
PowerEdge R750xs	370.0	0.6	1.0	2	104.6	182.5	21.9	9.5	16.6
SoC-Cluster (Kryo CPU)	672.0	1.3	0.2	2	437.4	76.5	27.0	32.4	5.7
SoC-Cluster (Adreno GPU)	387.0	193.8	X	2	37,500	X	27.0	2,777.8	X
SoC-Cluster (Hexagon DSP)	345.5	X	2,604.9	2	X	450,000	27.0	X	33,333.3

TABLE I
THEORETICAL COMPARISON BETWEEN SOC-CLUSTER AND CONVENTIONAL COTS EDGE SERVERS. "X" MEANS THAT THIS NUMERICAL OPERATION IS NOT SUPPORTED BY THE HARDWARE.

Our SoC-Cluster server (both its CPU and co-processors) have much higher computing capacity (either FP32 or INT8) per energy/size/weight than CPU servers.

Comparing satellite servers

	Throughput per Energy (TpE)			Throughput per Volume (TpV)			Throughput per Weight (TpW)		
	Power (watt)	GFLOPs per watt (FP32)	GINOPs per watt (INT8)	Volume (U)	GFLOPs per U (FP32)	GINOPs per U (INT8)	Weight (kg)	GFLOPs per kg (FP32)	GINOPs per kg (INT8)
Xeon 40-core CPU Server	276.3	0.8	0.5	1	208.3	130.4	18.8	11.1	6.9
NVIDIA A40 GPU Server	2,000.0	149.6	1,197.2	4	74,800	598,600	57.9	5,165.8	41,339.8
PowerEdge R350	95.0	0.5	0.9	1	49.3	85.4	13.6	3.6	6.3
PowerEdge R550	330.0	0.5	0.9	2	83.0	151.3	20.4	8.1	14.8
PowerEdge R750xs	370.0	0.6	1.0	2	104.6	182.5	21.9	9.5	16.6
SoC-Cluster (Kryo CPU)	672.0	1.3	0.2	2	437.4	76.5	27.0	32.4	5.7
SoC-Cluster (Adreno GPU)	387.0	193.8	X	2	37,500	X	27.0	2,777.8	X
SoC-Cluster (Hexagon DSP)	345.5	X	2,604.9	2	X	450,000	27.0	X	33,333.3

TABLE I

THEORETICAL COMPARISON BETWEEN SOC-CLUSTER AND CONVENTIONAL COTS EDGE SERVERS. "X" MEANS THAT THIS NUMERICAL OPERATION IS NOT SUPPORTED BY THE HARDWARE.

NVIDIA GPU has better capacity per size/weight. Yet, it is a monolithic server that (i) only accelerates domain-specific workloads, and (ii) has low reliability and flexibility.

Comparing satellite servers

北京郵電大學 Beijing University of Posts and Telecommunications

	Server Volume	Solar Panel Volume	Server Weight	Solar Panel Weight
Xeon 40-core CPU server	1	1.7–4.3	18.8	27.6–120.1
NVIDIA A40 GPU server	4	121.2–312.5	57.9	2,000.0–8,695.7
SoC-Cluster	2	4.1–10.5	27.0	67.2–292.2

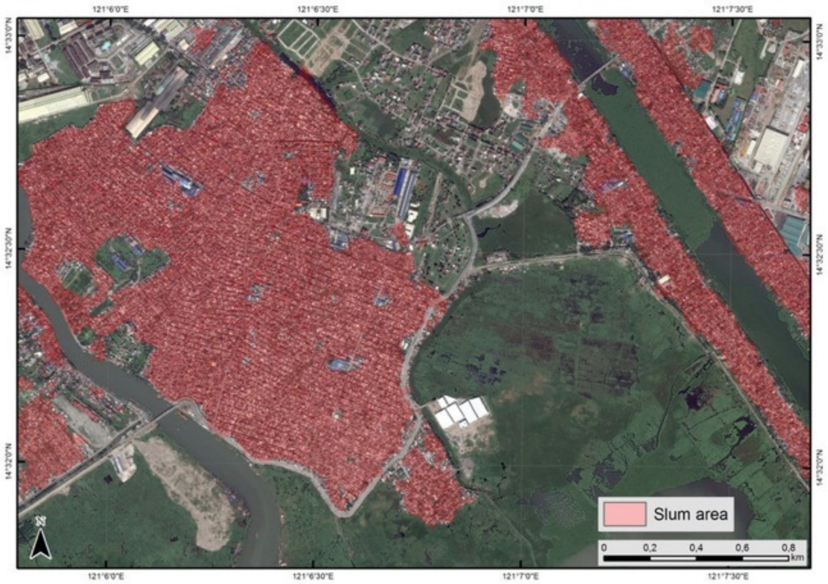
TABLE II

BOTTLENECK ANALYSIS OF IN-SPACE COMPUTING: THE SOLAR PANELS DEMANDED TO PROVIDE ENOUGH POWER IS MUCH HEAVIER AND LARGER THAN THE SERVER ITSELF. WE ASSUME THE AVERAGE SERVER UTILIZATION IS 50%.

A back-of-the-envelope calculation shows that energy is more likely to be the constraint than size/weight due to the reliance on a large solar panel.

Comparison with applications

- Video transcoding
  - Earth imagery pre-processing
- Deep learning inference
  - Object detection, segmentation, etc..



Comparison with applications

- Video encoding
  - Software: Ffmpeg & LiTr[1].
  - Datasets: 6 videos randomly picked from vbench[2]
- Deep learning inference
  - Software: TVM@Intel CPU; TensorRT@NVIDIA GPU; TFLite@SoC
  - Models: ResNet-50, ResNet152, YOLOv5x, BERT
- Alternative hardware
  - 40-core Intel Xeon Gold 5218R processor
  - NVIDIA A40 GPU.

Comparison with applications

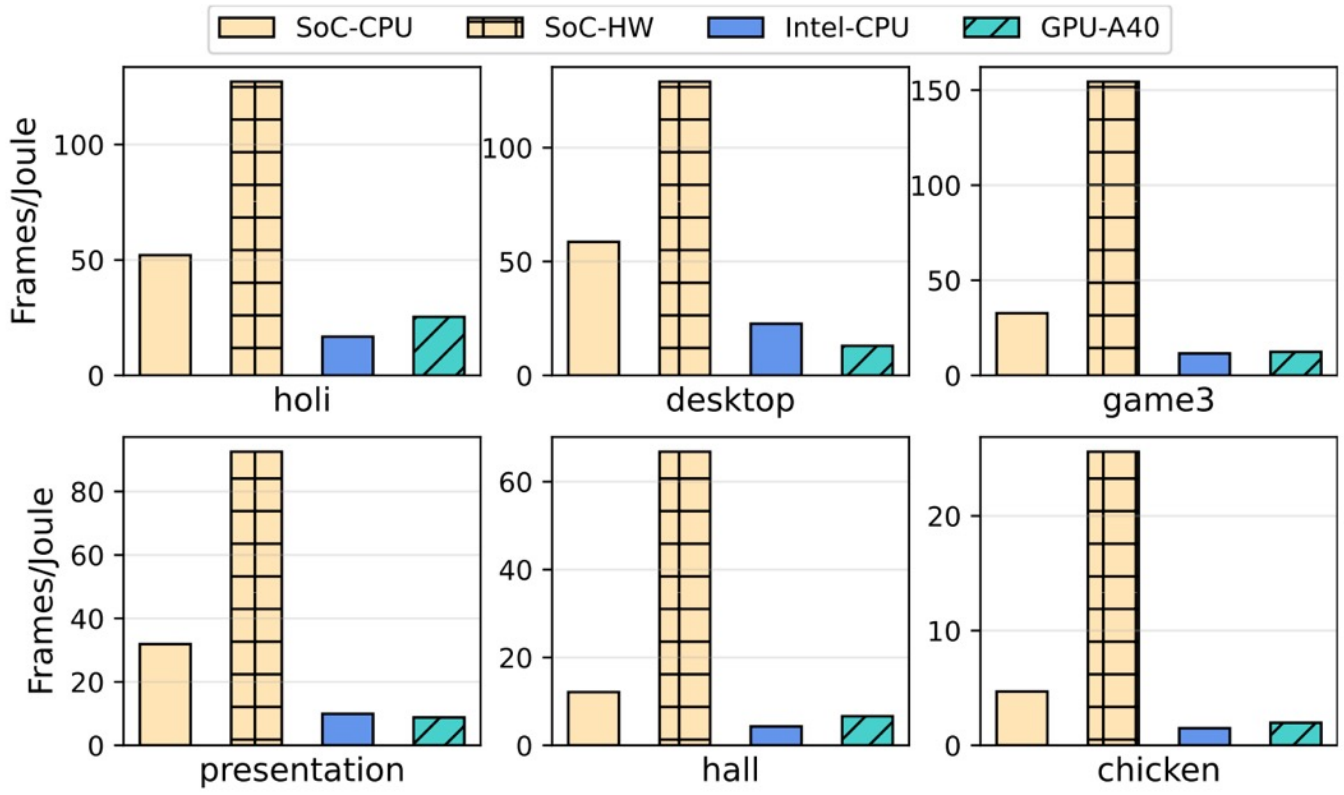


Fig. 4. Processed frames of per Joule (an indicator of TpE) on the video processing experiments. The 6 videos are randomly selected from a popular video benchmark [29].

- Our SoC servers can transcode 26–154 frames per Joule, which is 5.7×–17.1× higher than Intel CPU and 5.0×–13.0× higher than NVIDIA A40 GPU
  - Brings benefit even without using its hardware codec

Comparison with applications

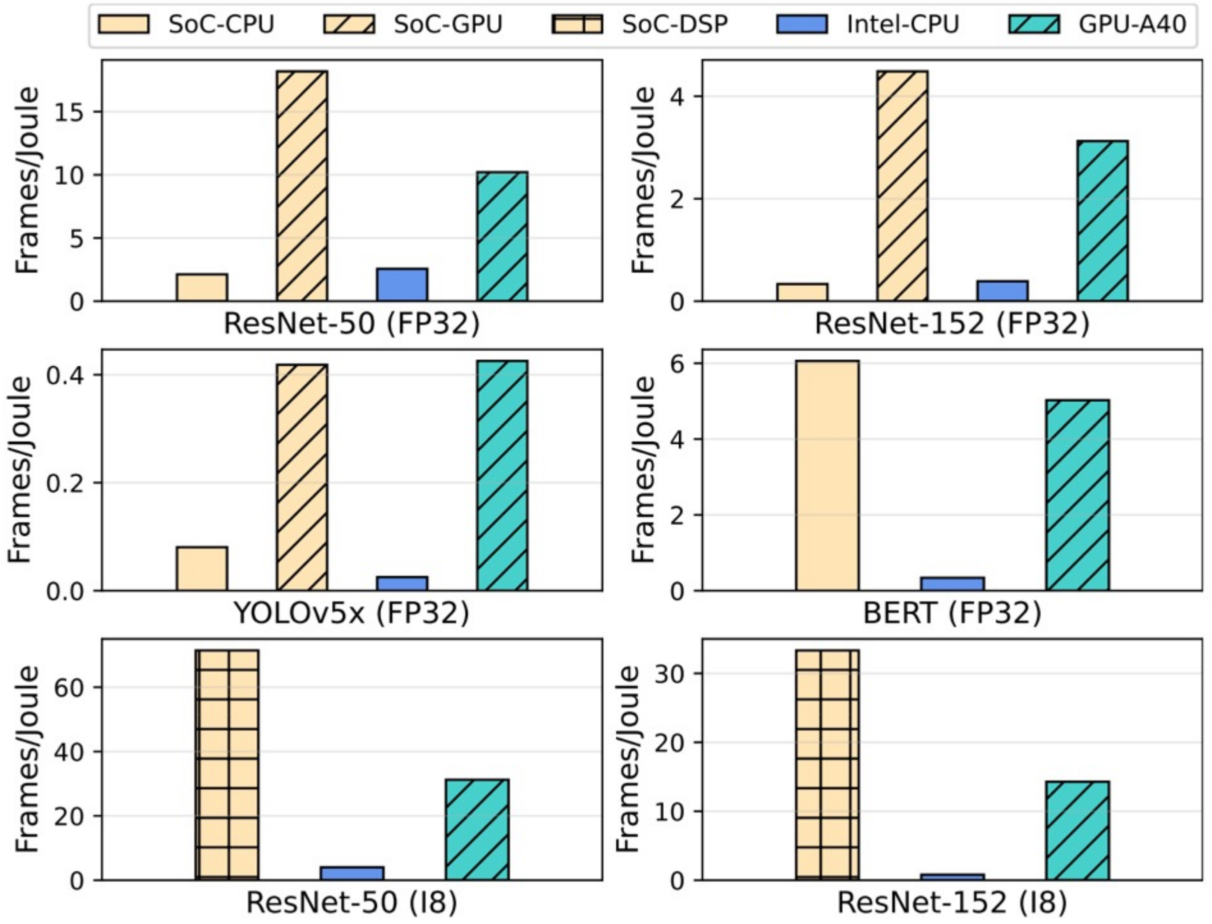


Fig. 3. Processed frames per Joule (an indicator of TpE) on the deep learning inference experiments. FP32: 32-bit floating point; I8: 8-bit integer.

- Running prediction with ResNet-50 model (FP32), SoC GPU can process 18.2 samples per Joule, which is 7× and 1.8× higher than Intel CPU and NVIDIA GPU, respectively.
- The energy efficiency of SoC DSP is even more significant, i.e., 2.3× higher than NVIDIA A40 GPU (with batch size 64).
- SoC can proportionally its energy efficiency with number of samples, while a monolithic NVIDIA GPU cannot

Comparing (to-be-)launched Satellite Servers

Name	Launched Time	Hardware Platform	Process	General-purpose cores	DSA capacity	Other specs
HPE Spaceborn Computer-2	2021.02 (ISS)	2x HPE Converged EL4000 Edge system 2x HPE ProLiant DL360 server	14nm	64*2 cores + 28*2 cores	-	2*1U, 2*14KG, 2*800W; 2*1U, 2*17KG, 2*800W
北邮一号	2023.01	2x RPI-4B 2x Atlas 200DK	28nm,	8 cores + 10 cores	26GFLOPS + 16TFLOPS/32TINOPS	Small enough, ~3KG, 13 W + 16 W
天智一号						<27KG
星测未来						
RUAG Space	xxxx	Lynx Single Board Computer ARM processor with > 30000 DMIPS		4 cores for single Board	-	25W
Exo-Space FeatherEdge	2023	Quad Cortex-A53 CPU	~10nm	4 cores	4 TOPS	
Our SoC Server	2023	50x SoC (Snapdragon 865 & Rockchip RK3588)	5nm – 8nm	400 cores	300TOPS (QS865: 62.5TFLOPS + 750TINOPS)	2U, 27KG, ~560W peak power

Takeaways

- Need for in-space computing is urgent
- A satellite-born server design: SoC-Cluster
  - Massive, low-power, sub-10 nm chips
  - Each SoC is heterogeneous itself (with GPU/NPU)
  - A decentralized architecture for reliability
- A set of experiments that demonstrates the advantages of SoC-Cluster over traditional servers
- We plan to launch the server into space in 2023!