Technical Trends in AI Infrastructure

Bhavesh Patel
Dell Technologies

REGIONAL SDC 24
BY Developers FOR Developers
APRIL 24, AUSTIN, TX

A SNIA. Event

Agenda

- About Me
- Accelerator Trends
- AI Infrastructure Trends
- What to consider?

About me

- Technologist and Sr. Distinguished Engineer
- 20 years at Dell Technologies
- Current focus:
  - Accelerator Strategy
  - AI Solution Strategy
- Ultra-runner

Compute for AI

CPU - SISD

- Complex control logic
- High programmability
- High-90s % of workloads and algorithms
- Low compute density

GPU - SIMD

- Excels at vectored floating point
- High compute density
- Hurt by branches or exceptions – “if” statements.
- Floating point data type

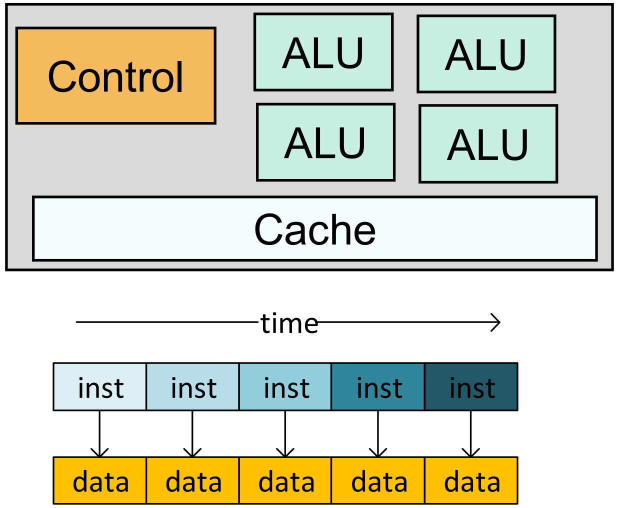
FPGA - MIMD

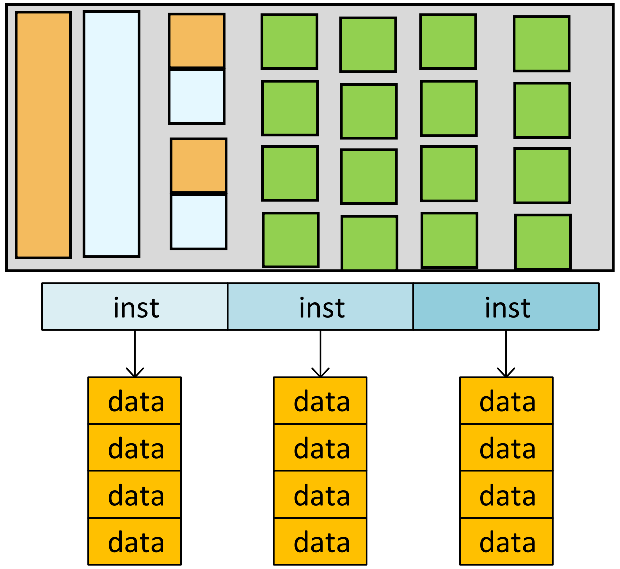
- Data-flow, pipeline oriented and/or vectored operation
- Very nimble at the bit-level
- Excellent streaming with IO devices

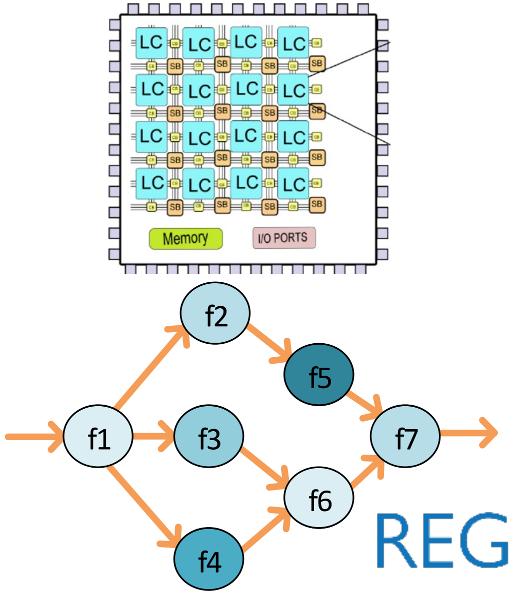
Domain Specific Accelerators

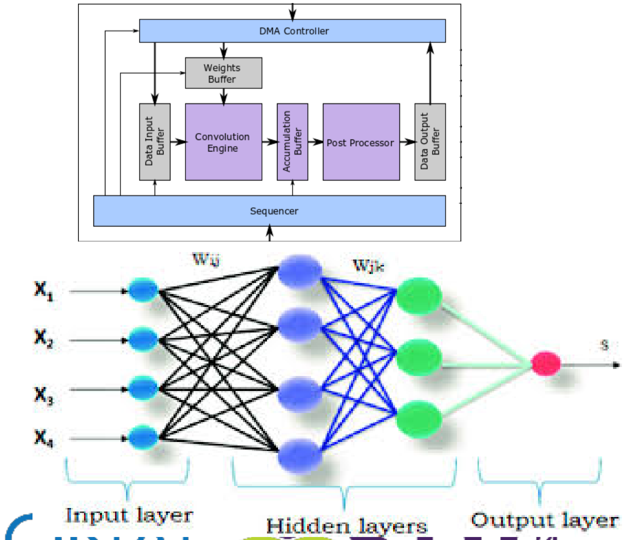
- Optimized for Matrix multiplications.
- Distributed high speed memory
- Targeting specific workloads

- Remains the core center of computation
- Remains focused on a subset of high-performance problems
- Optimized for specific use cases.
- Fully optimized for specific workloads.
- Better compute utilization

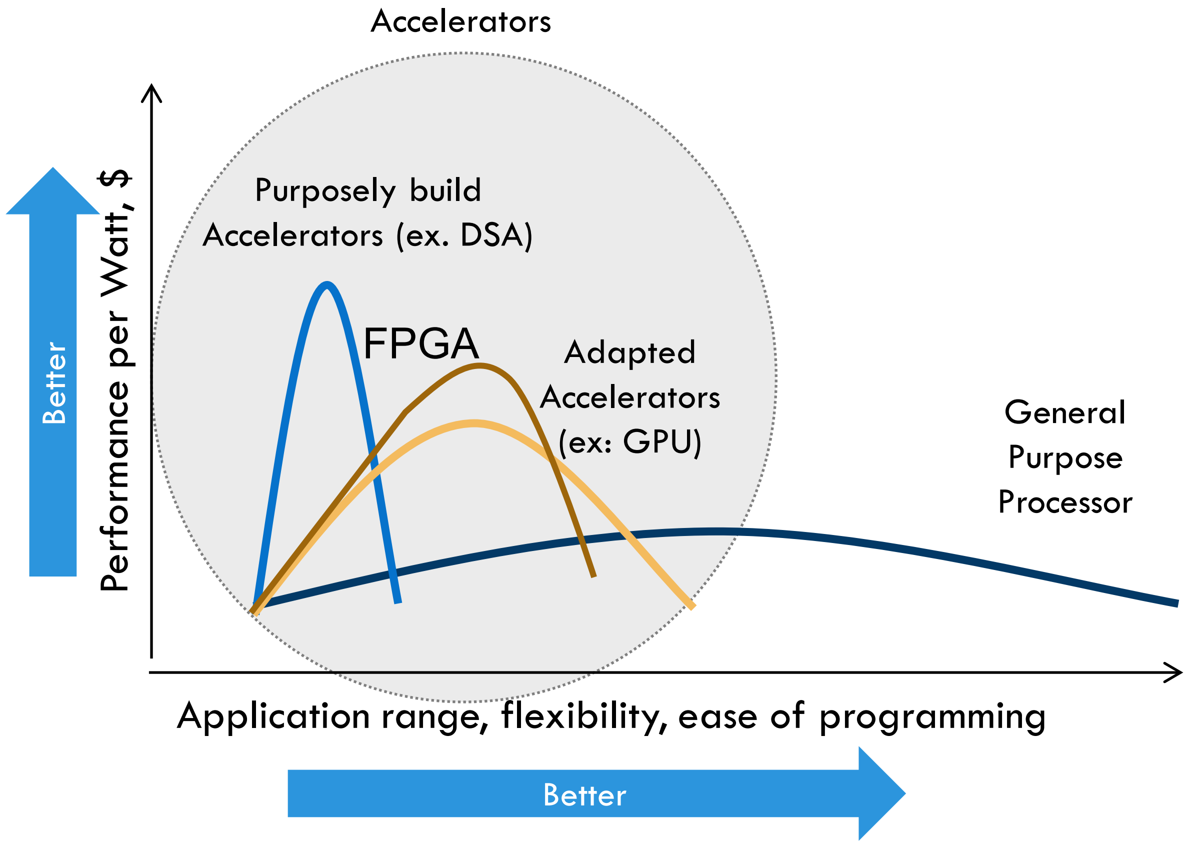








Processing Landscape



Several specialized accelerators are emerging, aiming to provide

- Faster Fine-tuning and inference
- Better processing efficiency and
- Better solution cost (CAPEX + OPEX)

vs. general purpose processors.

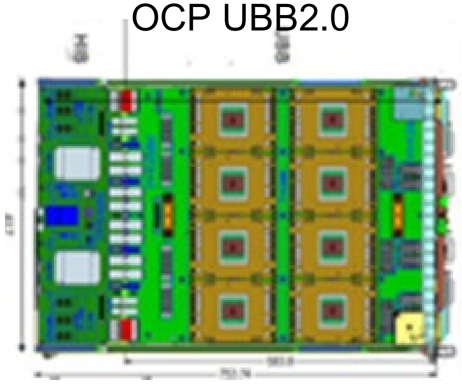
Building AI System requires tradeoff: flexibility, app range vs processing efficiency

REGIONAL SDC 24

Silicon Trends for AI Compute

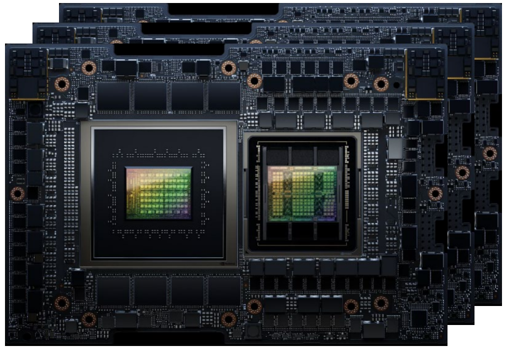
Training

OCP UBB2.0



Wafer scale Custom

Coherent CPU-GPU



6KW -8KW

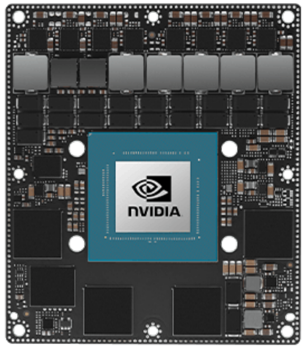
20KW

>50KW

Inferencing

M.2

Embedded



Low Power PCIe CEM

Med Power PCIe CEM

High Power PCIe CEM

5-10W

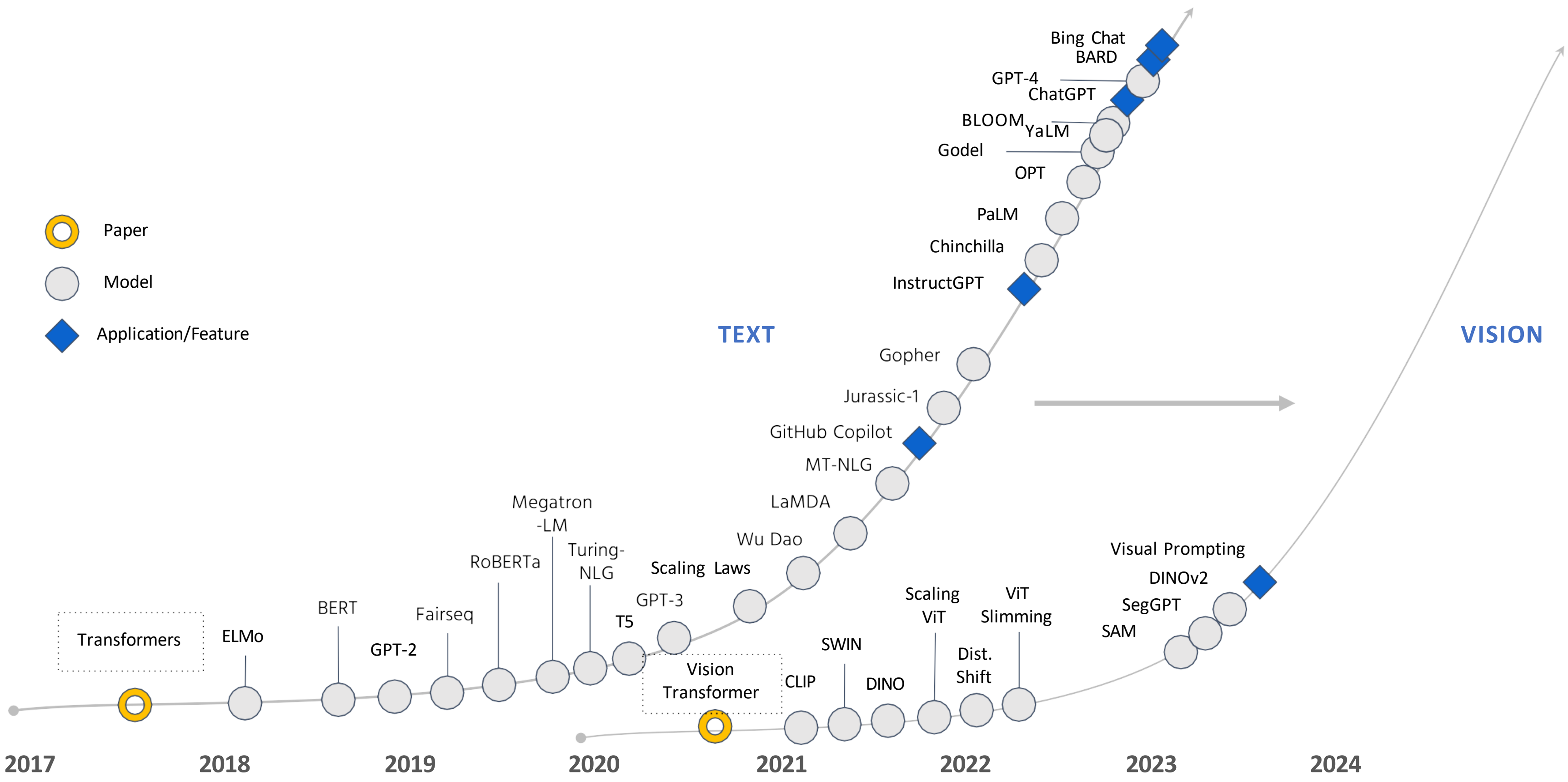
15W – 40W

75W

150W

300W -600W

LLM Models in Text & Vision Space



Andrew Ng

REGIONAL SDC 24

Large Language Model Compute Demand

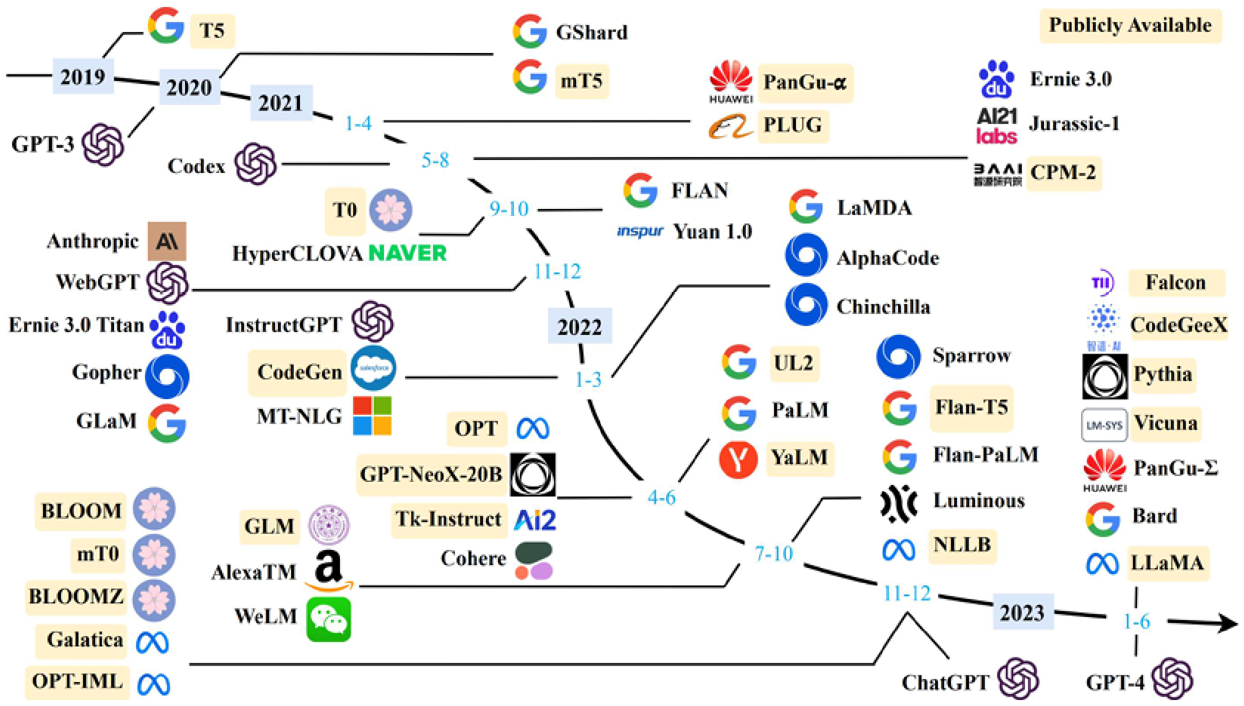


Figure 1

*image credit: Wayne Xin Zhao, et.al, “A Survey of Large Language Models”

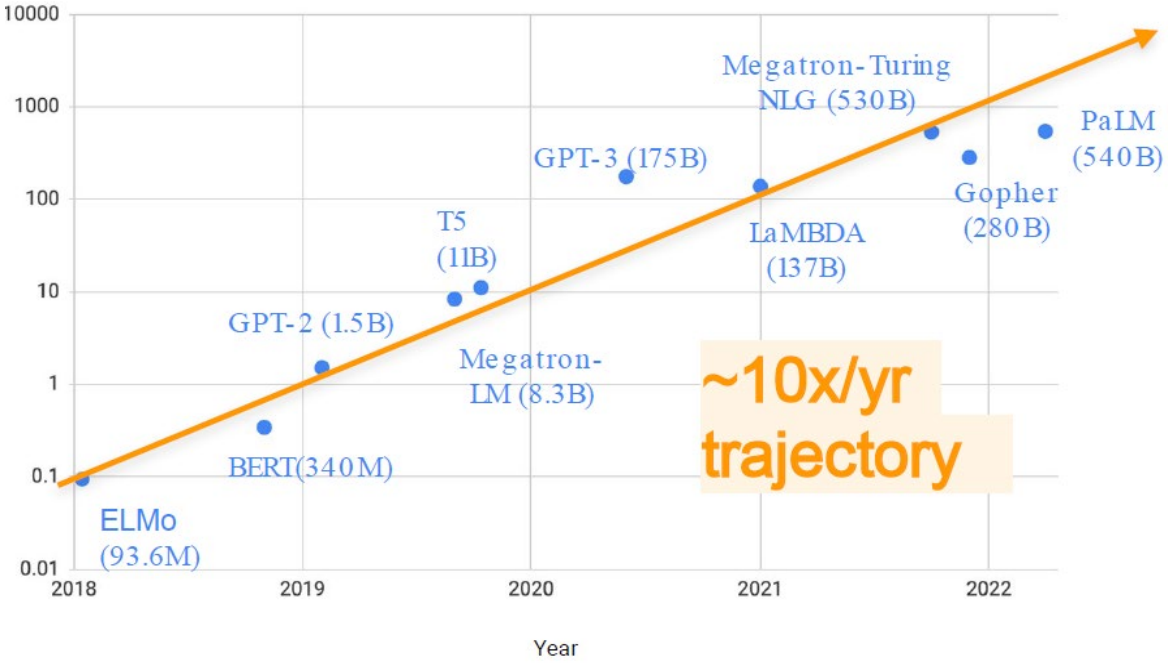


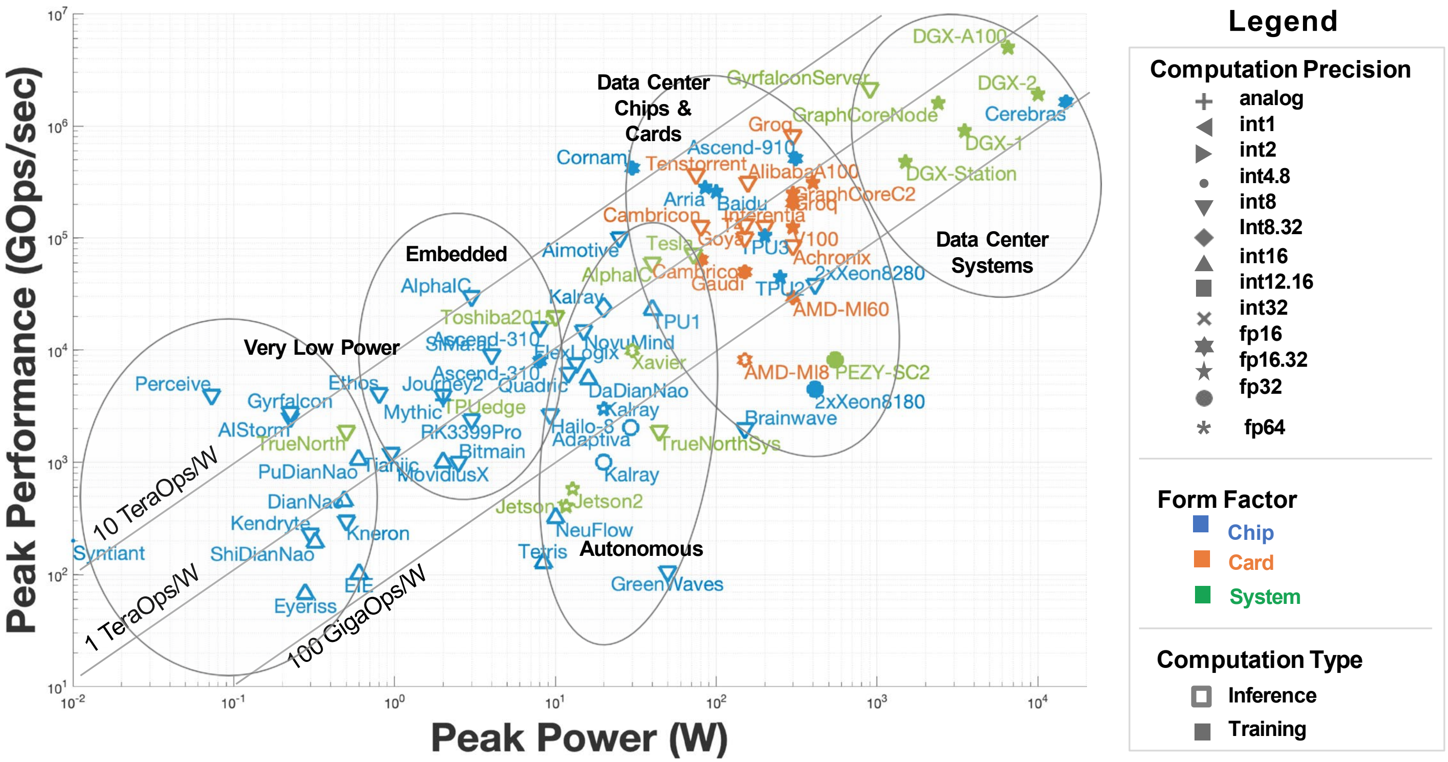
Figure 2

*image credit: Amin Vahdat at AI Hardware and Edge AI Summit 2023

LLM size has been increasing exponentially over last 5 years and will continue, which enforces both performance and cost for the hardware to run those models.

REGIONAL SDC 24

Machine Learning Accelerators



Peak performance vs. power scatter plot of publicly announced AI accelerators and processors.

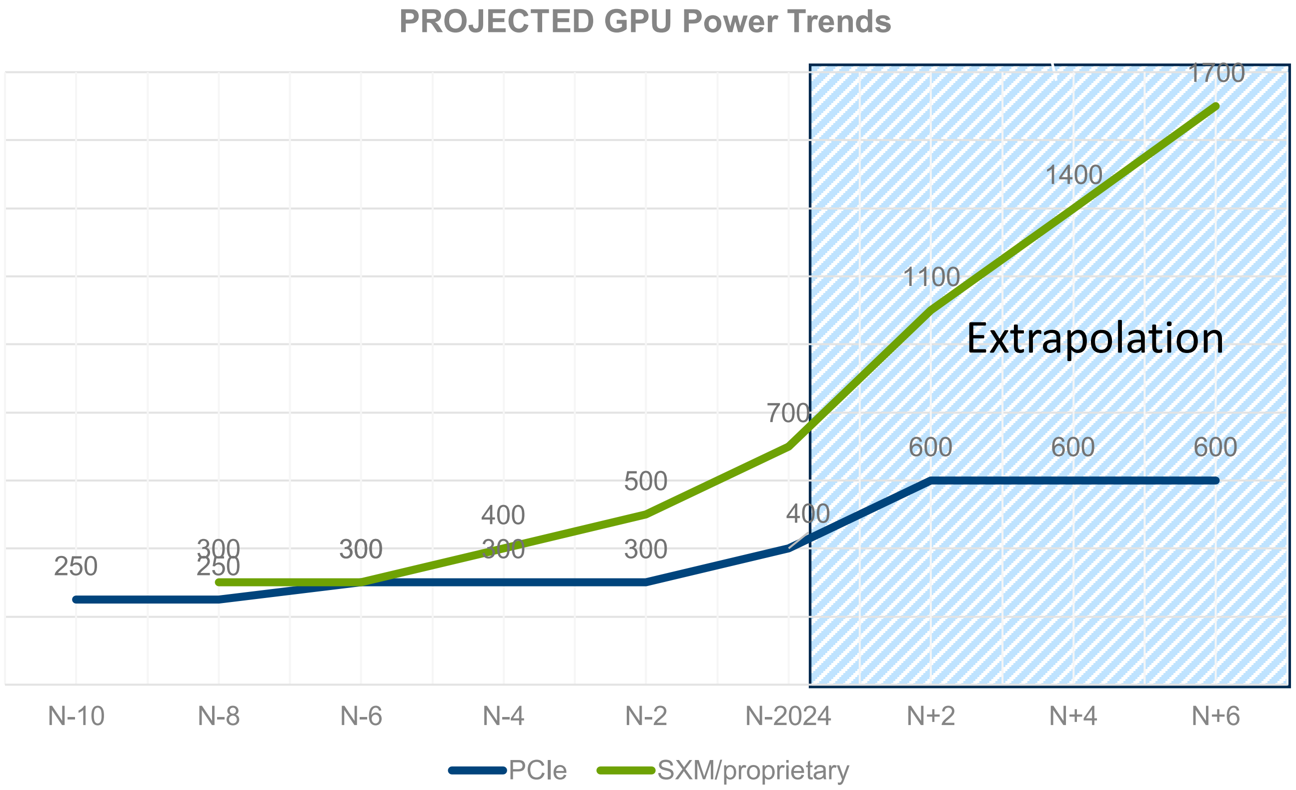
Source:
Survey of Machine Learning Accelerators

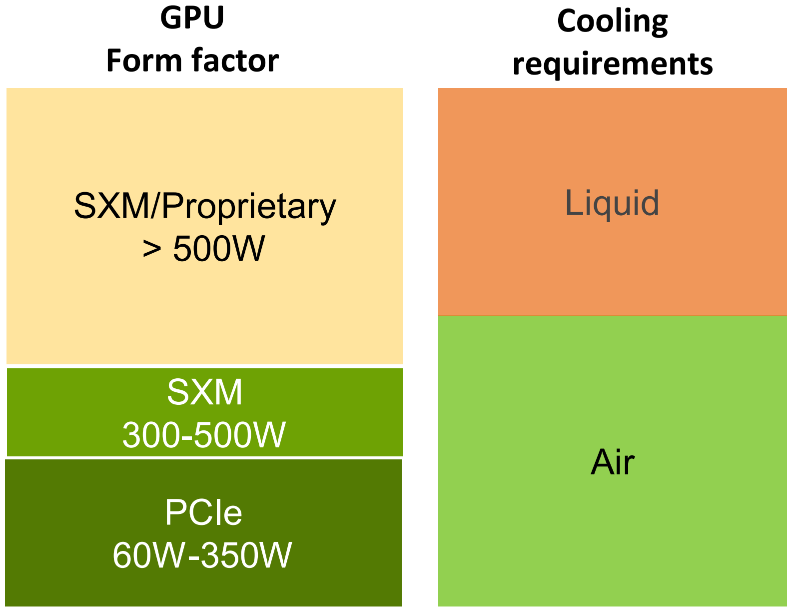
Albert Reuther, Peter Michaleas, Michael Jones, Vijay Gadepally, Siddharth Samsi, and Jeremy Kepner
MIT Lincoln Laboratory Supercomputing Center
Lexington, MA, USA
freuther,pmichaleas,michael.jones,vijayg,sid,kepnerg@ll.mit.edu

AI Compute Power Trends

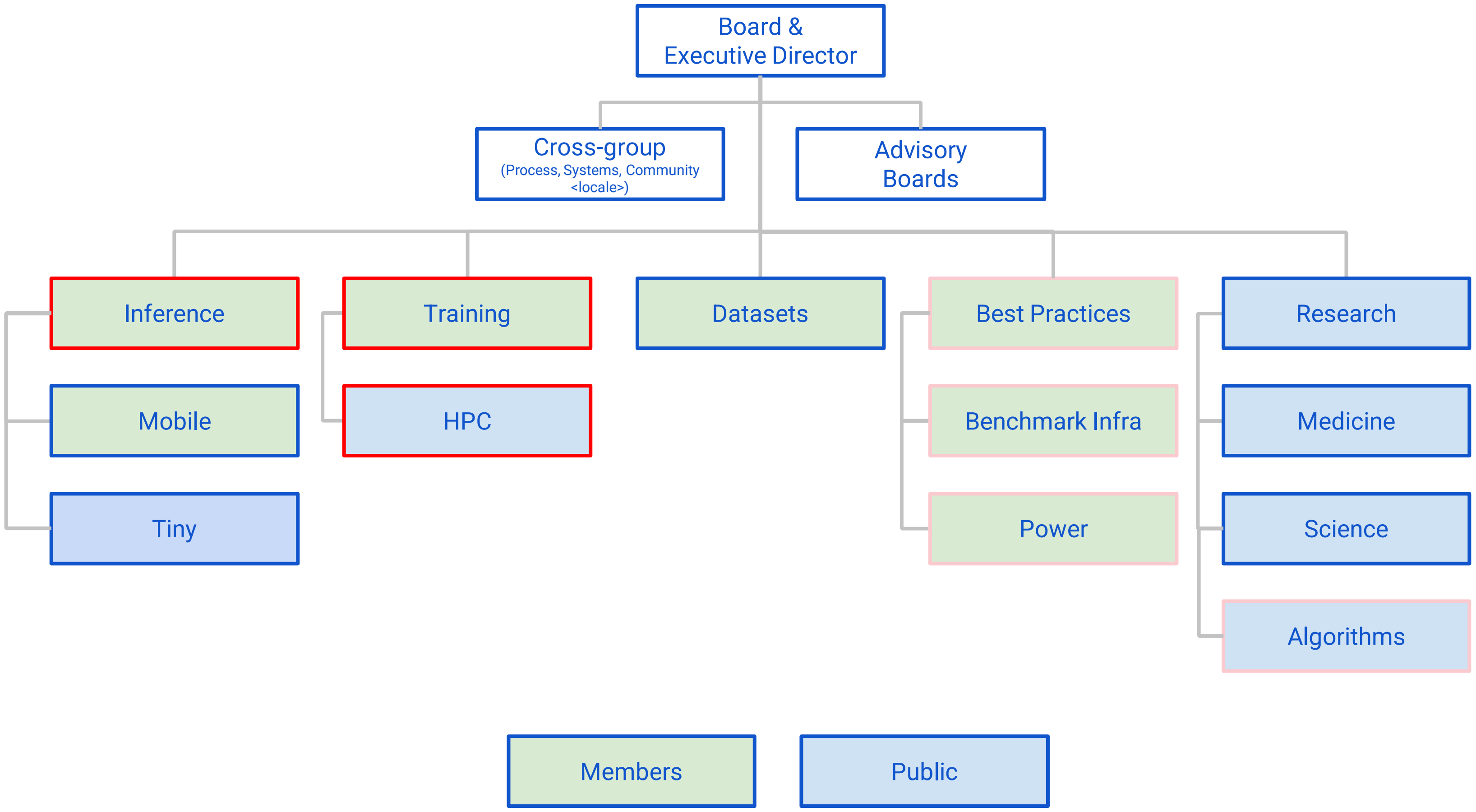
Rising GPU power trends impact solution design

Performance drives up consumption and cooling





Machine Learning Benchmarks: MLCommon Org Chart



REGIONAL SDC 24

Models

Model	Reference code	Framework	Dataset	Task
resnet50-v1.5	vision/classification_and_detection	tensorflow, onnx, tvm, ncnn	imagenet2012	Image classification
retinanet 800x800	vision/classification_and_detection	pytorch, onnx	openimages resized to 800x800	Object detection
bert	language/bert	tensorflow, pytorch, onnx	squad-1.1	Question answering
dlrm-v2	recommendation/dlrm_v2	pytorch	Multihot Criteo Terabyte	Recommendation
3d-unet	vision/medical_imaging/3d-unet-kits19	pytorch, tensorflow, onnx	KiTS19	Medical Image segmentation
rnnt	speech_recognition/rnnt	pytorch	OpenSLR LibriSpeech Corpus	Speech to text
gpt-j	language/gpt-j	pytorch	CNN-Daily Mail	Text Summarization
stable-diffusion-xl	text_to_image	pytorch	COCO 2014	Text to Image
llama2-70b	language/llama2-70b	pytorch	OpenOrca	Q&A Chatbot

Inference Categories

Data Center

Scenario	Query Generation	Duration	Samples/query	Latency Constraint	Tail Latency	Performance Metric
Server	LoadGen sends new queries to the SUT according to a Poisson distribution	600 seconds	1	Benchmark specific	99%*	Maximum Poisson throughput parameter supported
Offline	LoadGen sends all samples to the SUT at start in a single query	1 query and 600 seconds	At least 24,576	None	N/A	Measured throughput

Edge

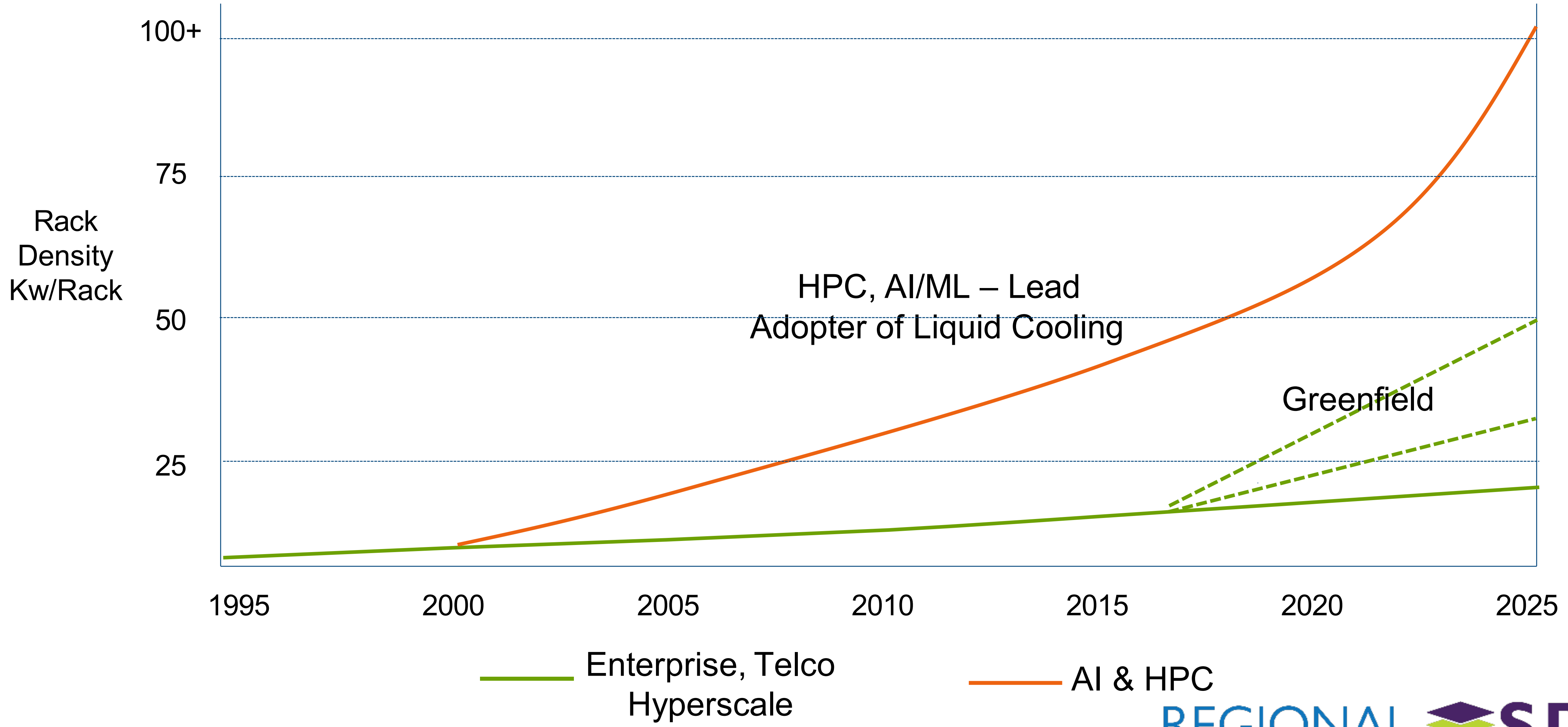
Scenario	Query Generation	Duration	Samples/query	Latency Constraint	Tail Latency	Performance Metric
Single stream	LoadGen sends next query as soon as SUT completes the previous query	600 seconds	1	None	90%*	90%-ile early-stopping latency estimate
Offline	LoadGen sends all samples to the SUT at start in a single query	1 query and 600 seconds	At least 24,576	None	N/A	Measured throughput
Multistream	Loadgen sends next query, as soon as SUT completes the previous query	600 seconds	8	None	99%*	99%-ile early-stopping latency estimate

Training Categories

The closed division models and quality targets are:

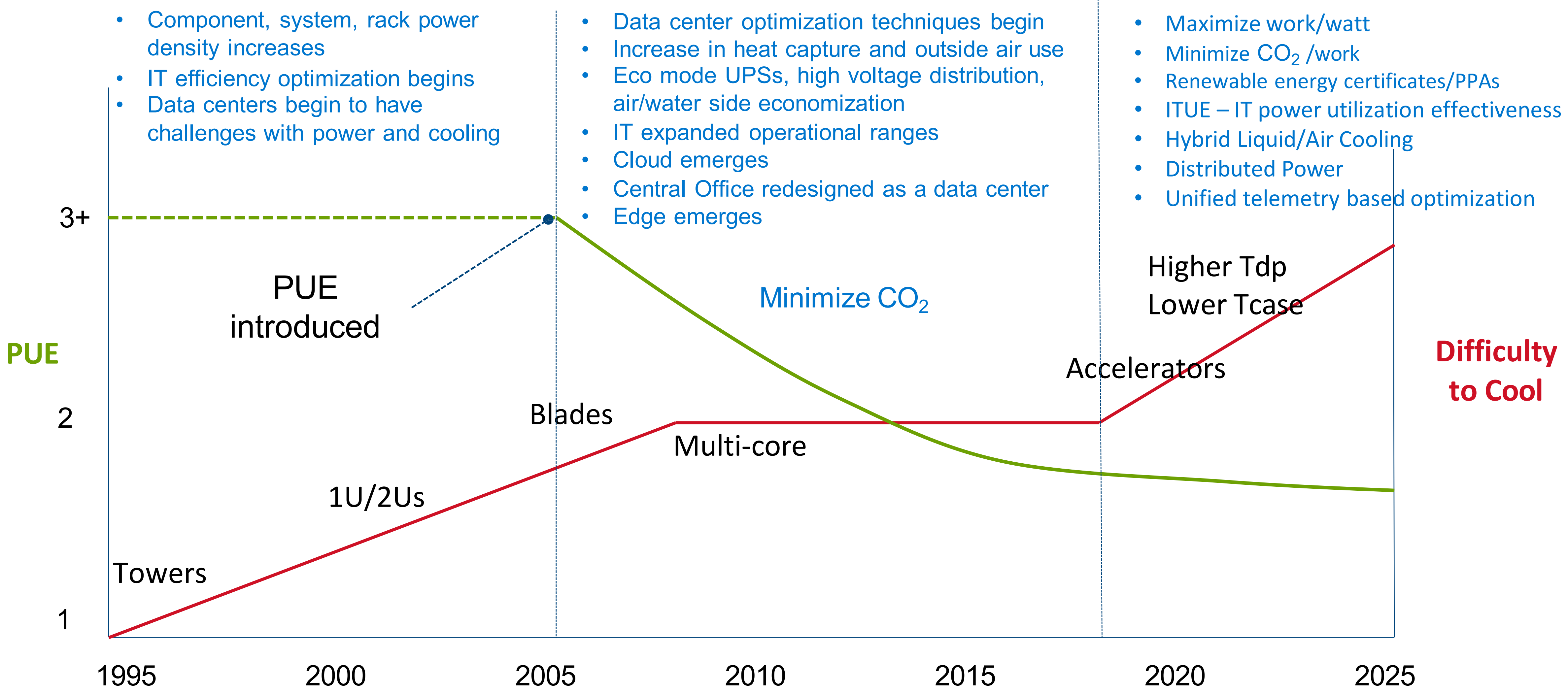
Area	Problem	Model	Target
Vision	Image classification	ResNet-50 v1.5	75.90% classification
	Image segmentation (medical)	U-Net3D	0.908 Mean DICE score
	Object detection (light weight)	SSD (RetinaNet)	34.0% mAP
	Object detection (heavy weight)	Mask R-CNN	0.377 Box min AP and 0.339 Mask min AP
	Text to image	Stable Diffusion v2.0	FID⇐90 and and CLIP>=0.15
Language	Speech recognition	RNN-T	0.058 Word Error Rate
	NLP	BERT	0.720 Mask-LM accuracy
	Large Language Model	GPT3	2.69 log perplexity
Commerce	Recommendation	DLRMv2 (DCNv2)	0.80275 AUC

Rack Power Density Trends

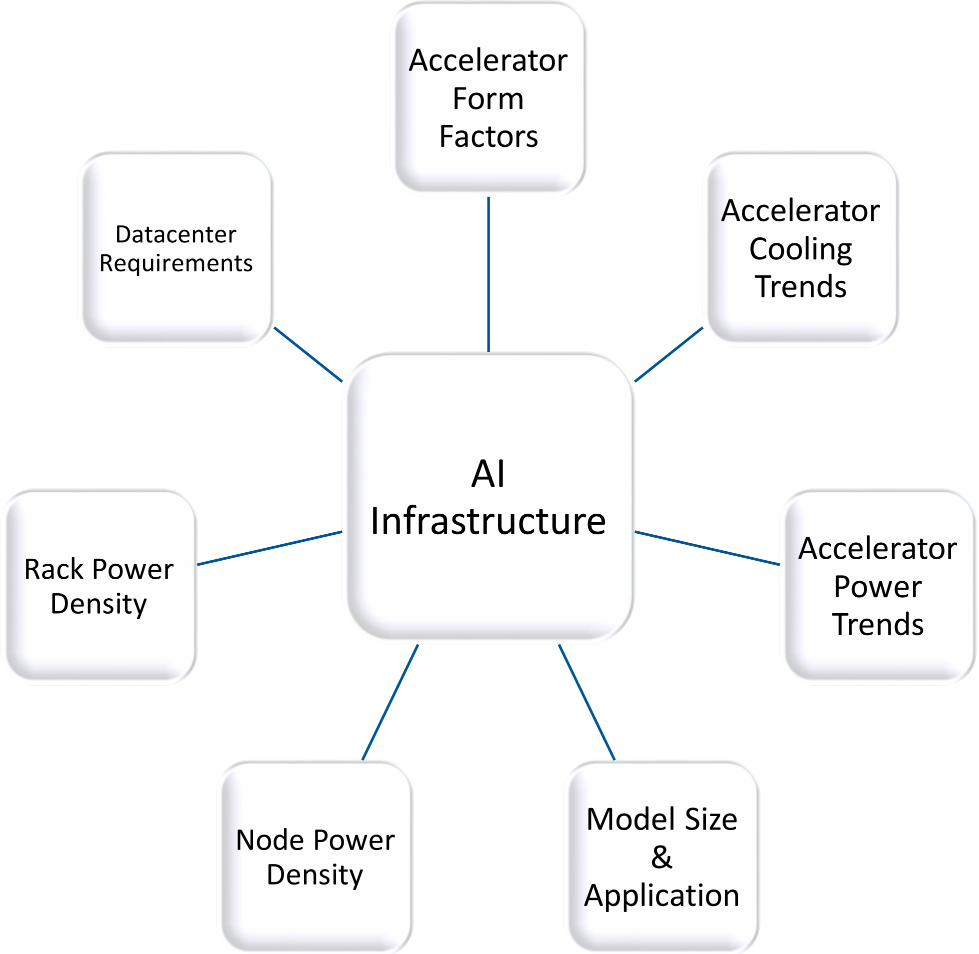


REGIONAL SDC 24

Datacenter Trends



Factors Impacting Infrastructure Decisions



REGIONAL SDC 24

Where is AI Infrastructure Trending?

- Moving from deployments using ‘individual’ compute nodes housing 2x – 8x accelerators to more ‘pod’ based deployment.
- For deploying GenAI applications and doing large language model training/fine-tuning or inferencing, the minimum pod size ranges from 64 – 1000x GPUs.
- This is impacting how we approach deployment of AI platforms i.e. the design essentially starts at the datacenter i.e. a Top down approach.
- The power per rack is increasing to meet AI compute demands & any prediction might be off by 20-30%.
- Fabric is becoming the core when designing AI Infrastructure because it has a direct impact on performance and scaling.

THANK YOU

Please take a moment to rate this session.

REGIONAL SDC 24
BY Developers FOR Developers