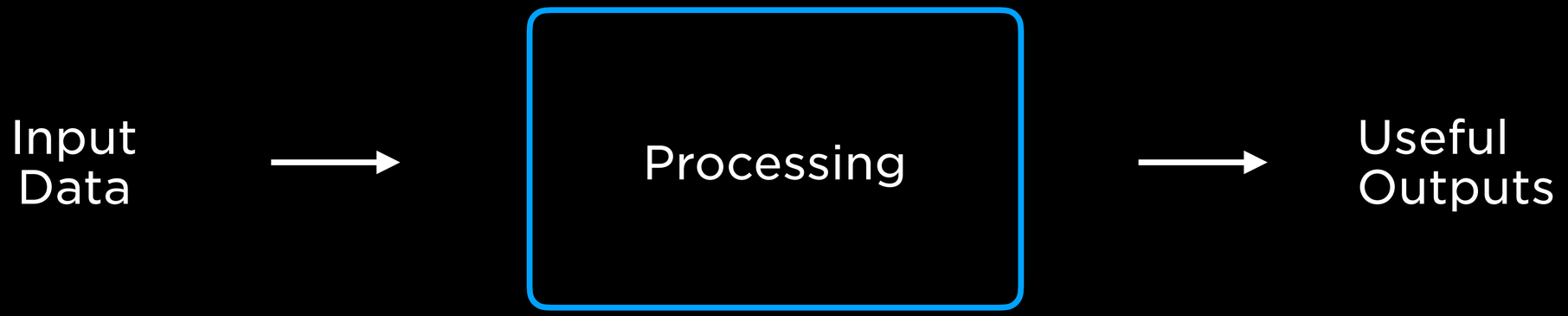
Beyond Compute:

Enabling AI Through System Integration

TESLA

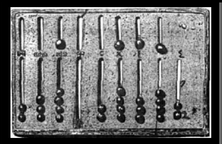
Computing



Computing

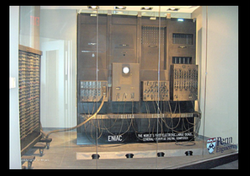
Input Data → Processing → Useful Outputs

Human Computer



Abacus

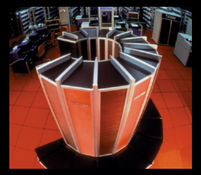
Diff Engine



ENIAC

Calculator

Personal Computer



Cray-1

Laptop Computer

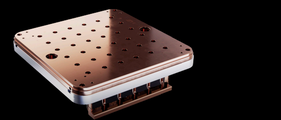
Consoles

Smart Phone Computer

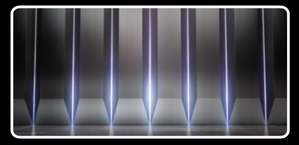
Datacenters

IoT Computer

FSD Computer



Training Server

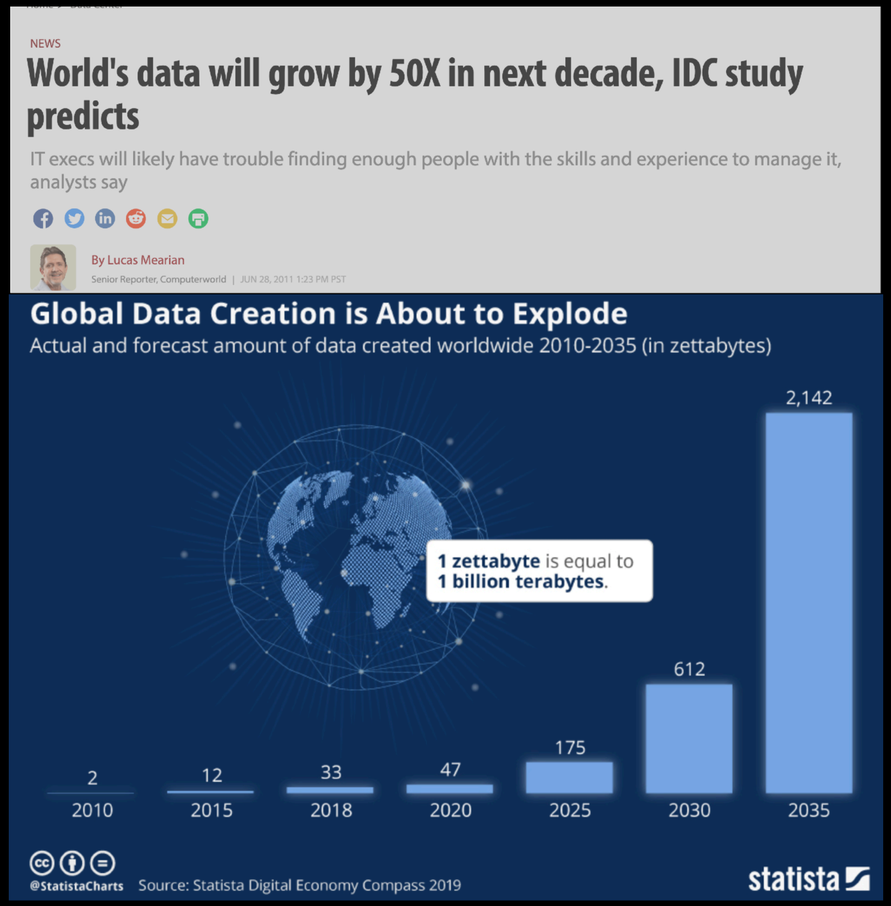


Training Datacenters

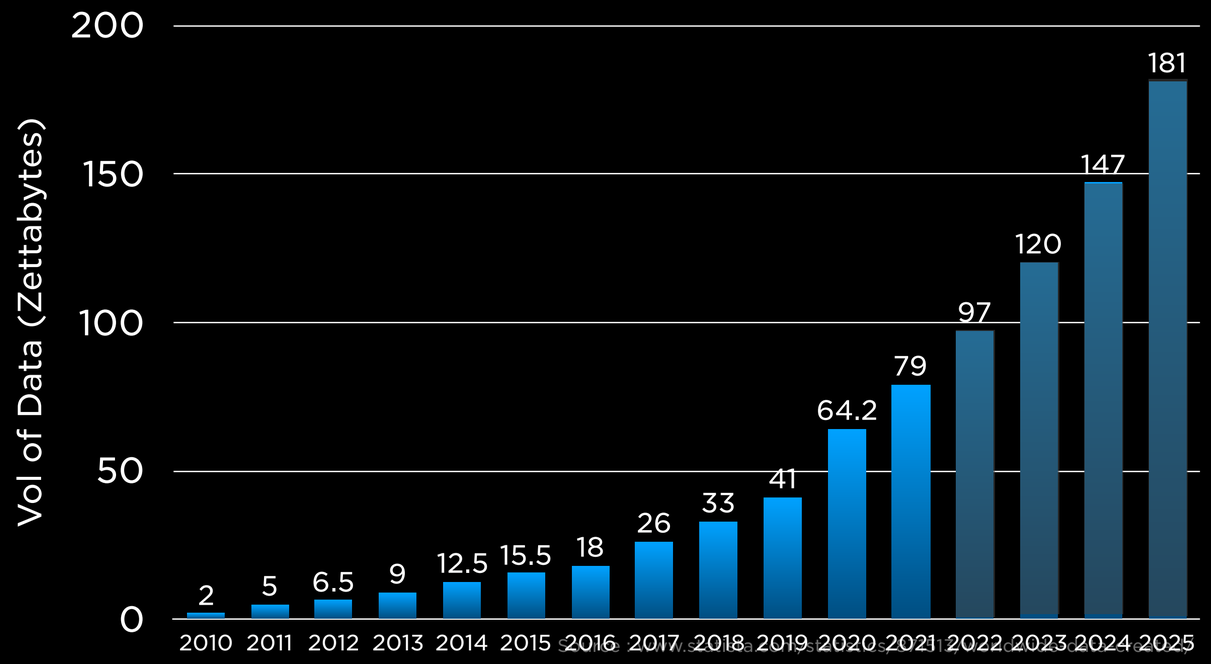
Source: Google Images

Big Data - Data Explosion

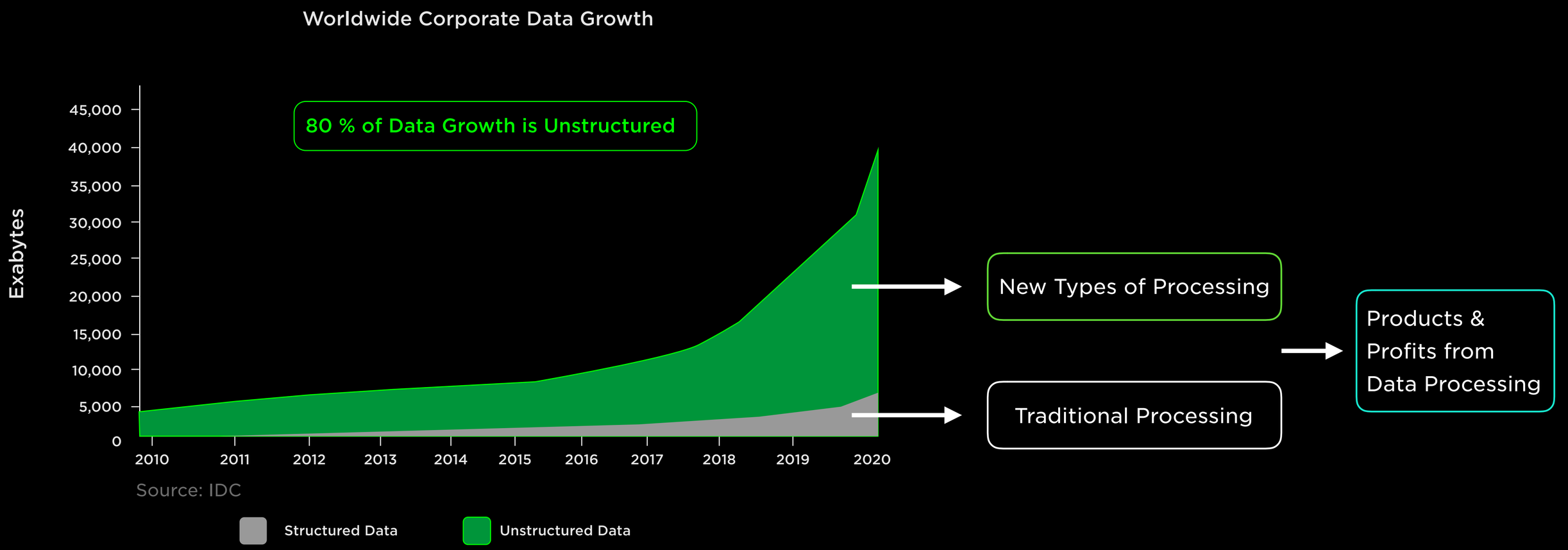
Projections In 2011



2021 Reality + New Projections

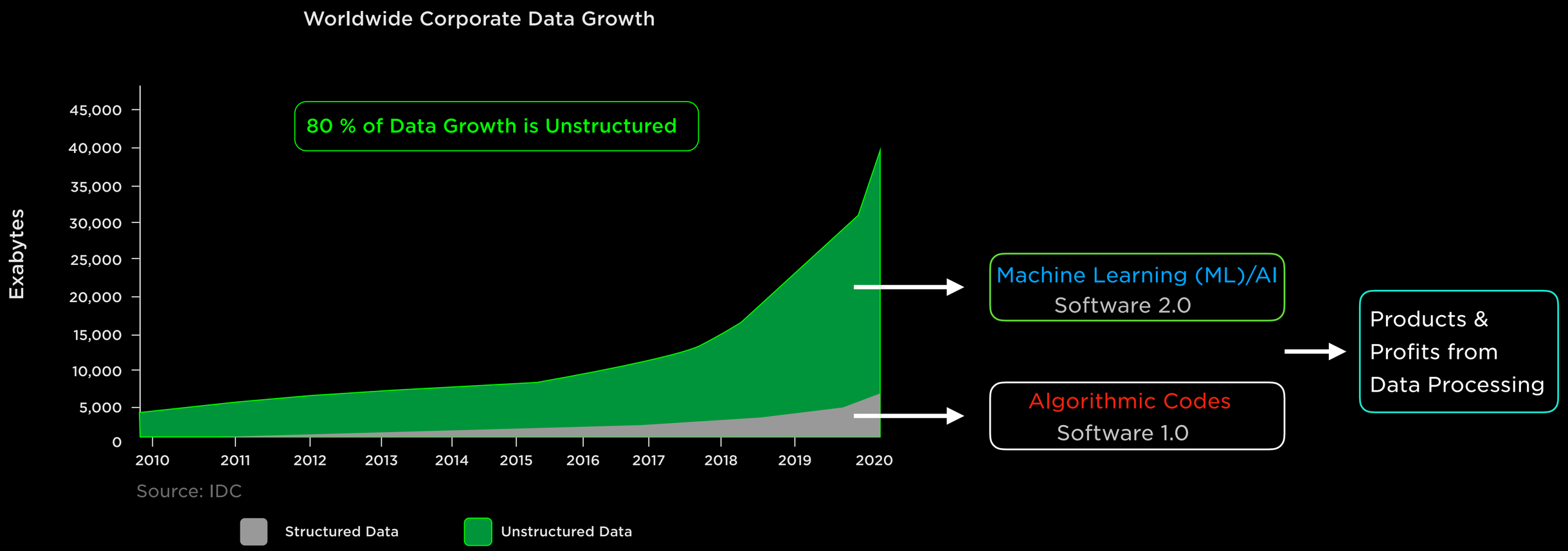


Corporate Data Type Changes



More than 80% data is unstructured in nature

Corporate Data Type Changes

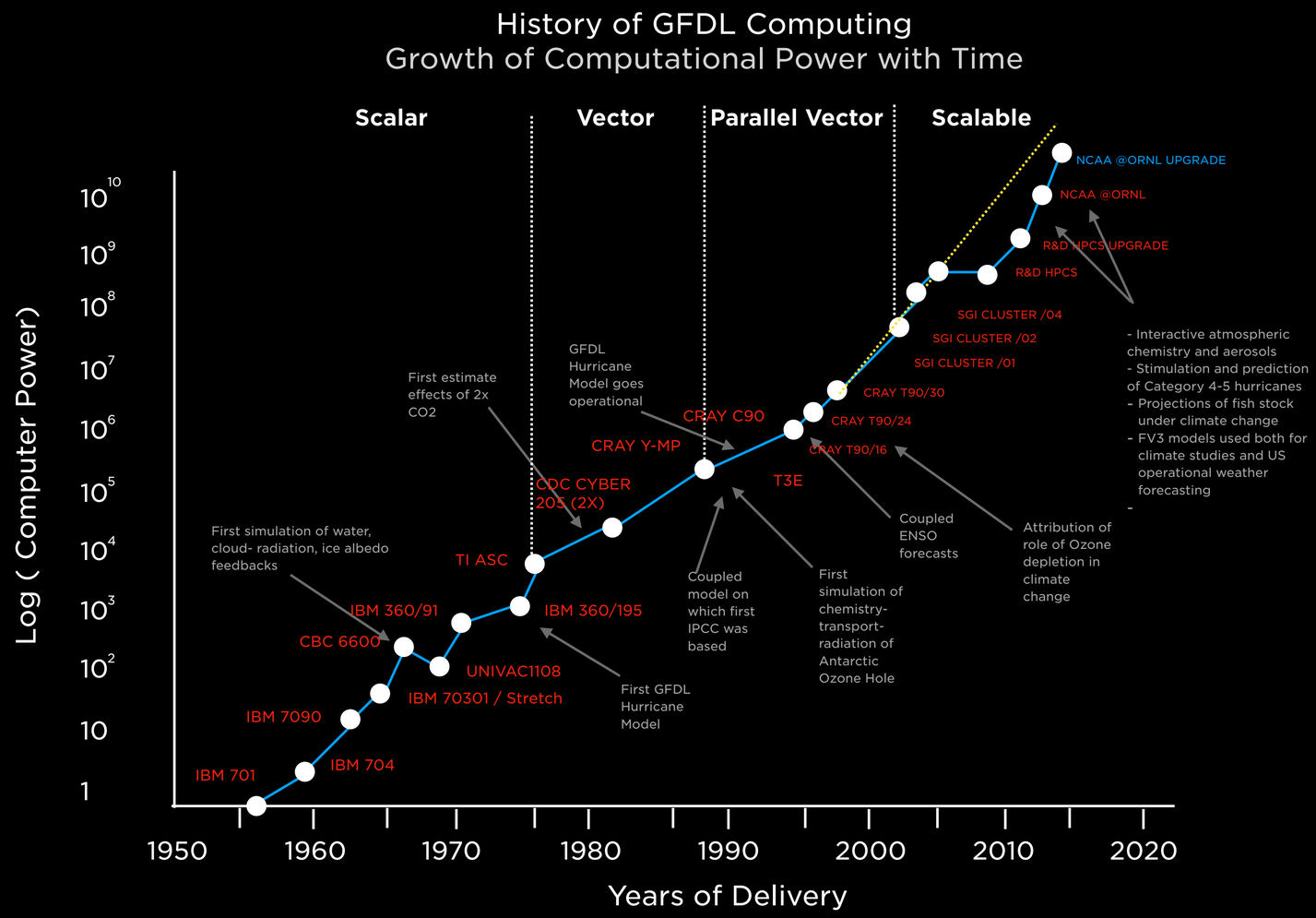


More than 80% data is unstructured in nature

Climate Super-Computing Architectures Over Time

"The history of numerical weather prediction and climate simulation is almost exactly coincident with the history of digital computing itself."

-V. Balaji , Climbing down Charney's ladder: Machine Learning and the post-Dennard era of computational climate science



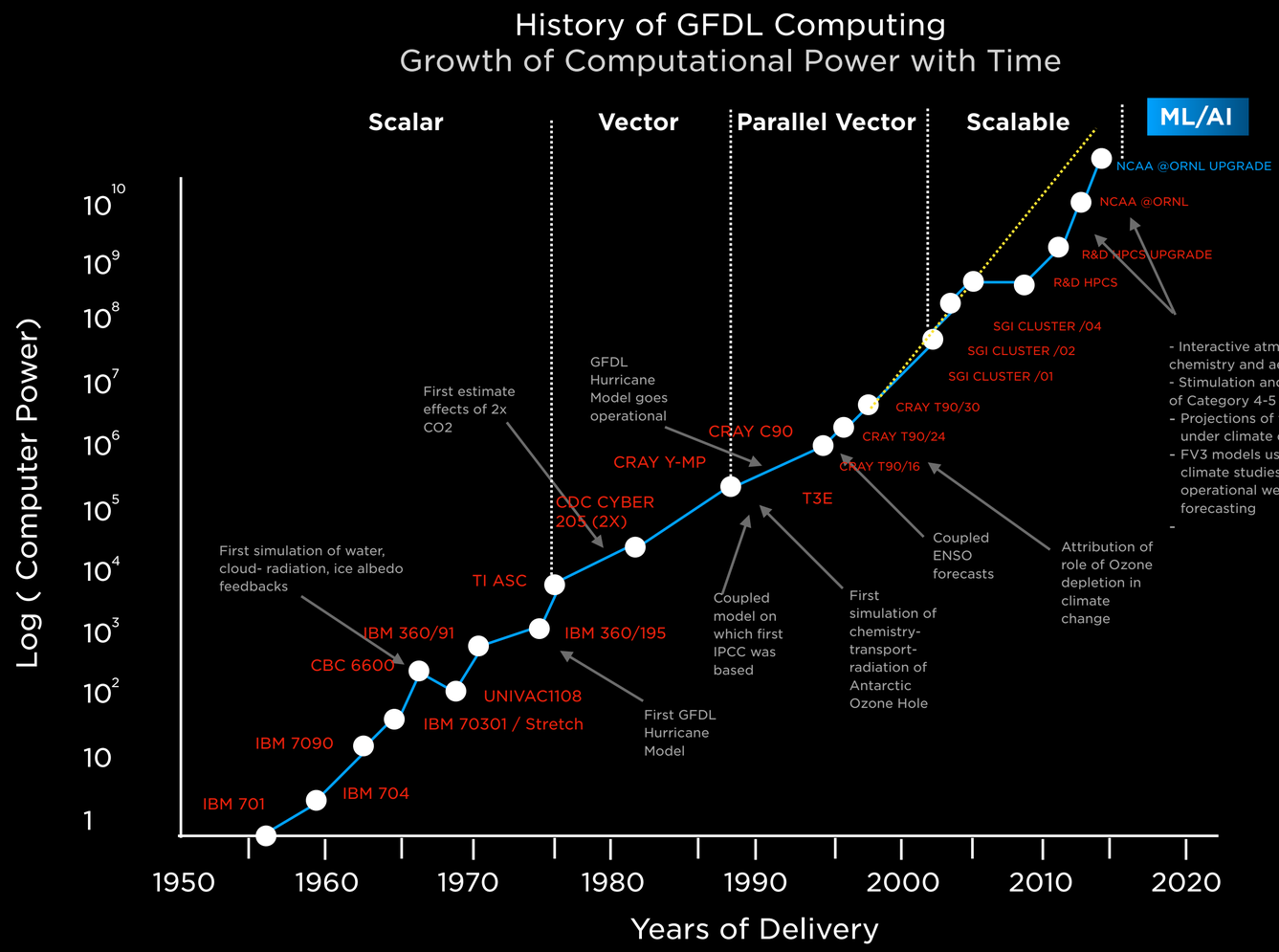
History of computational power at the NOAA Geophysical Fluid Dynamics Laboratory. Computational power is measured in aggregate floating point operations per second

Source: V Balaji, Princeton-NOAA :Climbing down Charney's ladder: Machine Learning and the post-Dennard era of computational climate science

Climate Super-Computing Architectures Over Time

"... this represents a sea change in computational Earth system science that rivals the von Neumann revolution."

- On the Increase of Machine Learning techniques in climate computing
  As quoted in Machine Learning and the post-Denard era of computational climate science



History of computational power at the NOAA Geophysical Fluid Dynamics Laboratory.
Computational power is measured in aggregate floating point operations per second

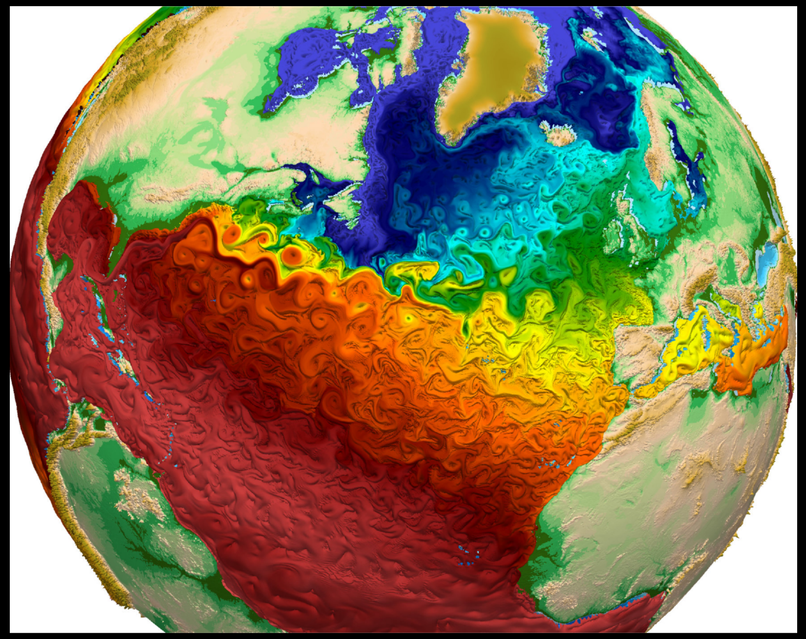
Source: V Balaji, Princeton-NOAA :Climbing down Charney's ladder: Machine Learning and the post-Dennard era of computational climate science

Traditional Data - New Processing Methods

Climate AI, a pioneer in applying artificial intelligence to climate risk modeling, today announced its team has solved a critical weather forecasting challenge. Leveraging advances in AI to improve weather and climate forecasts.

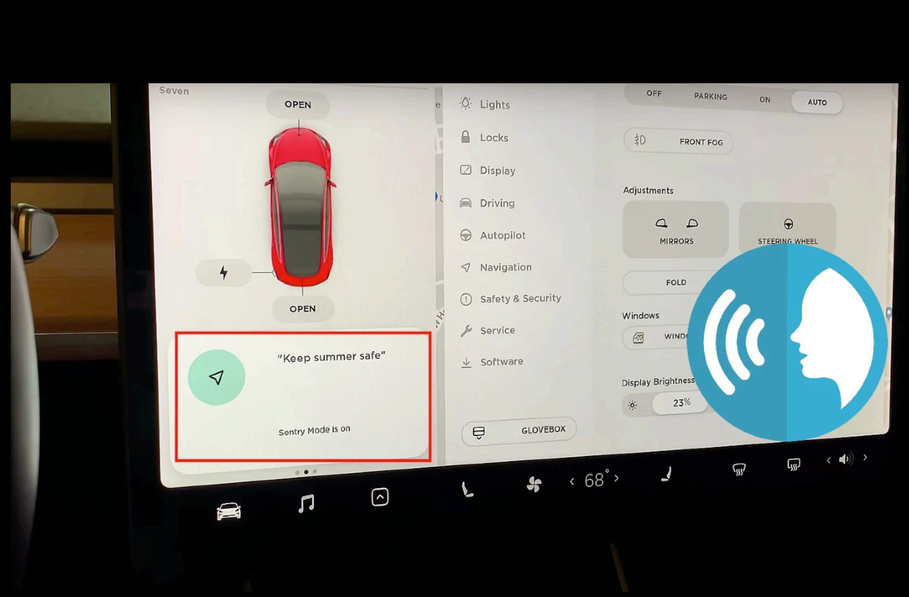
"Artificial intelligence and machine learning breakthroughs are changing weather forecasting, and resource-heavy regional weather models might soon be completely replaced by machine learning approaches." Dr. Stephan Rasp, Lead Data Scientist Climate AI

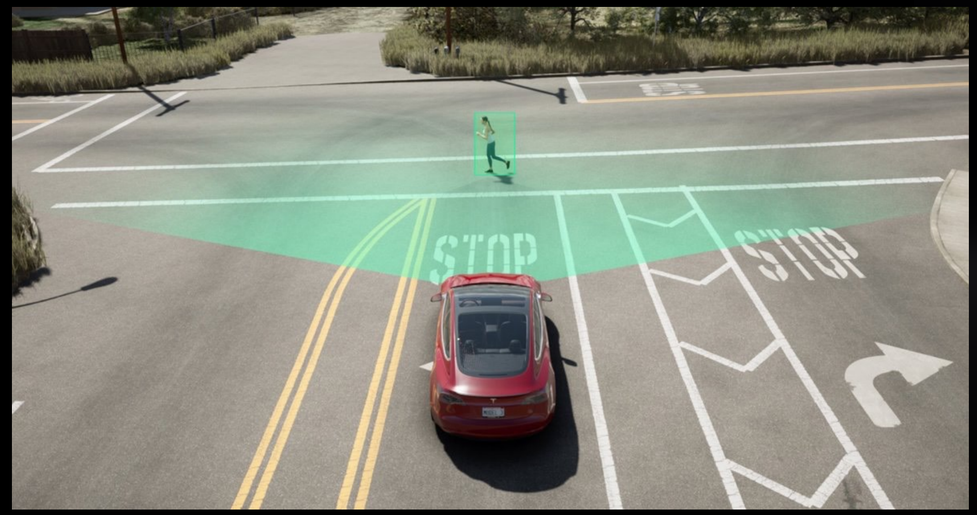
Source: Climate AI



Source:Los Almos National Lab

Real World Data





Only Machine Learning techniques can enable these

Exploding ML Use Cases

New Data Types

Traditional Data

Real World Data

....

.

Computing Architecture Categories

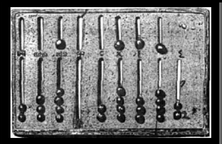
Input Data → Processing → Useful Outputs

Accelerators : Strictly Structured

Algorithmic Computers: Semi-Structured

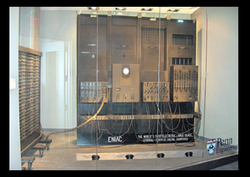
Learning Computers : Unstructured/Any type

Human Computer



Abacus

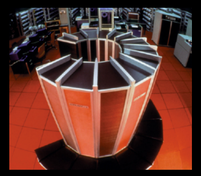
Diff Engine



ENIAC

Calculator

Personal Computer



Cray-1

Laptop Computer

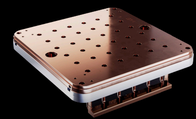
Consoles

Smart Phone Computer

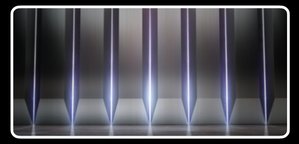
Datacenters

IoT Computer

FSD Computer



Training Server



Training Datacenters

Processing for Learning : AI - ML - DL

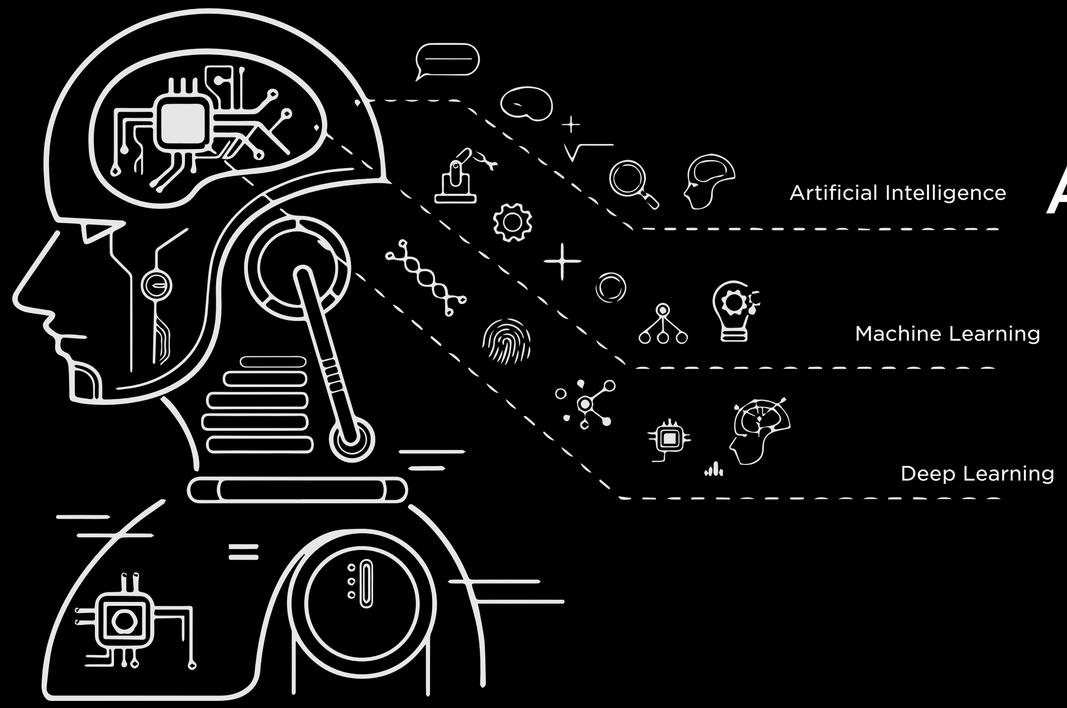


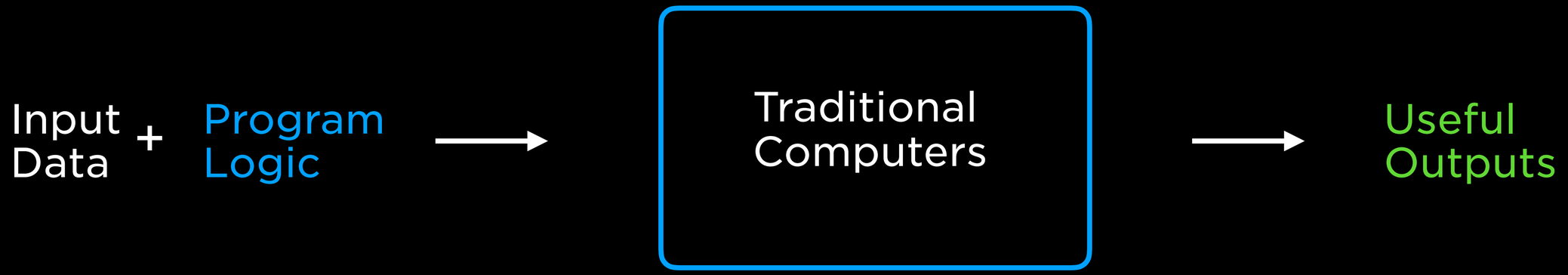
Image Courtesy: Buffaloboy

AI - Tasks requiring near human intelligence in real world settings

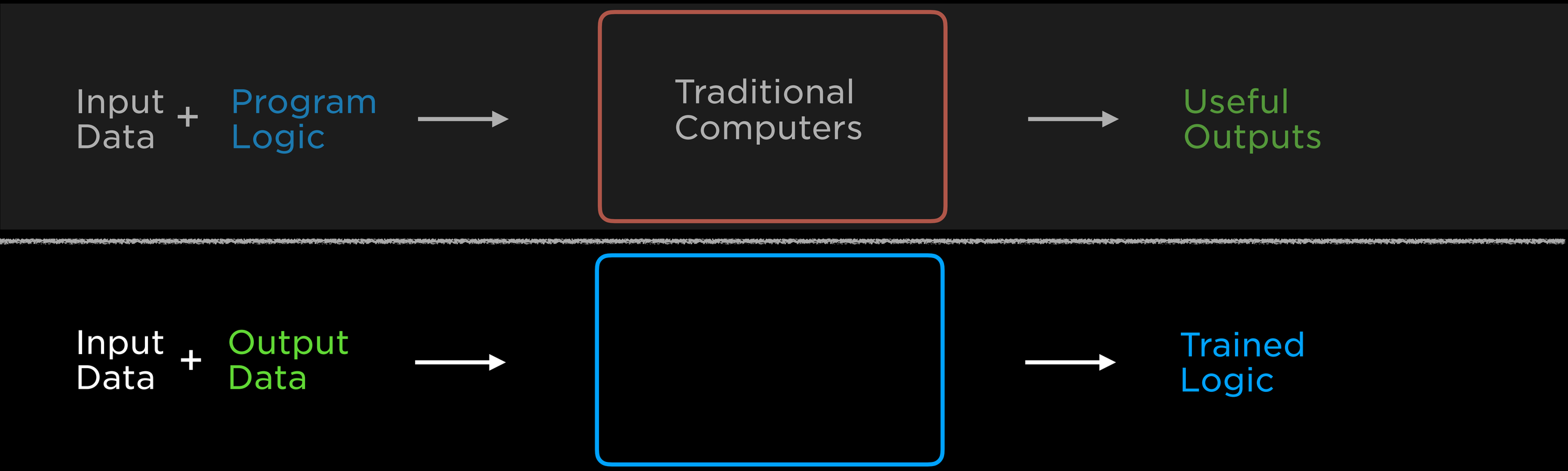
ML - Subset of AI for specific tasks by learning from data & making predictions

DL - Subset of ML using Deep neural network architectures

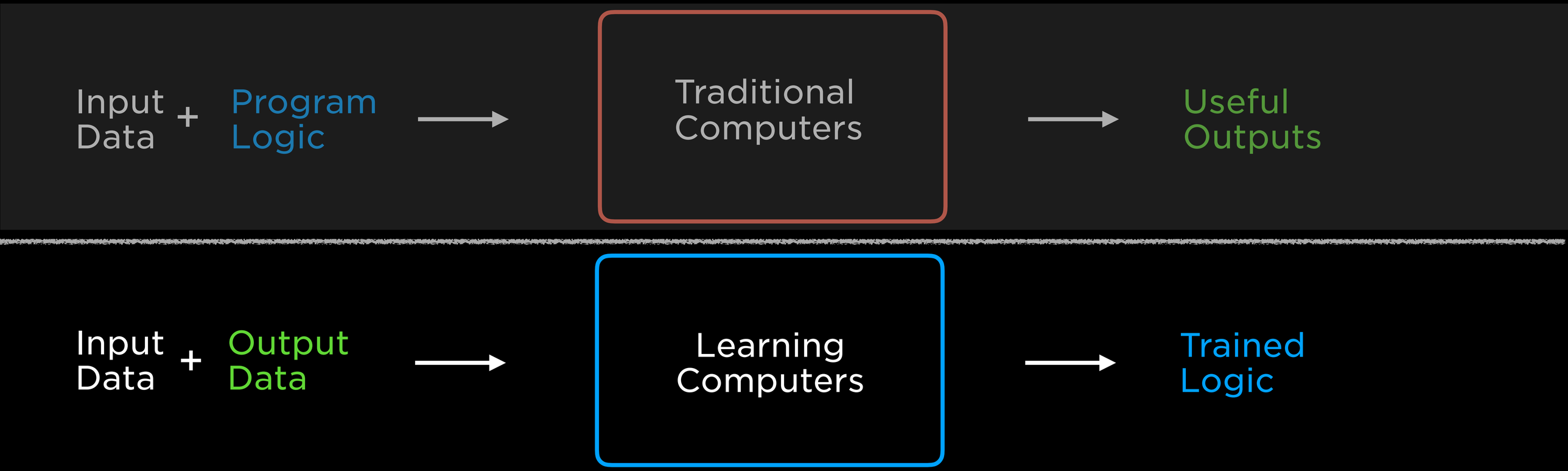
Why Do We Need a Different Compute Platforms?



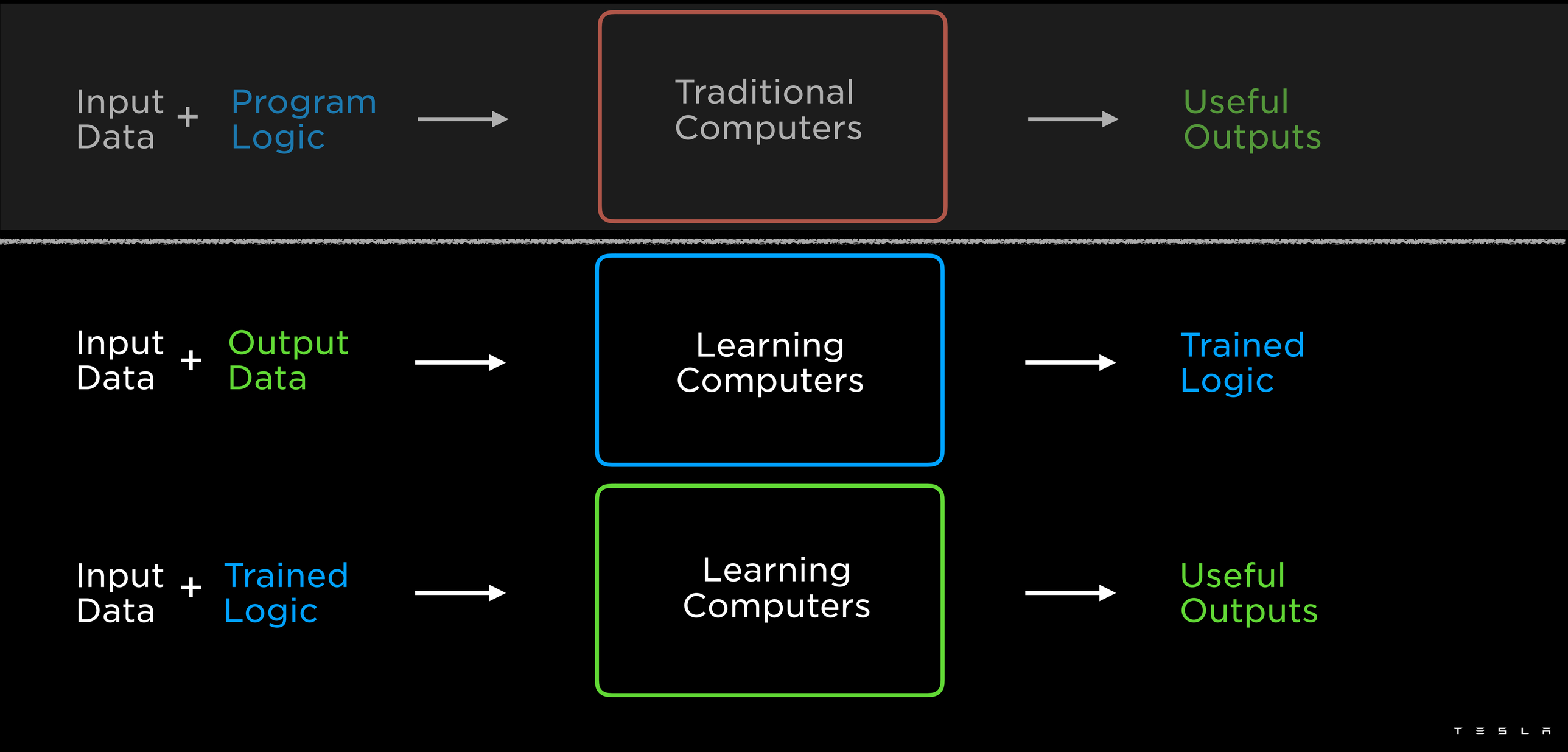
Why Do We Need a Different Compute Platforms?



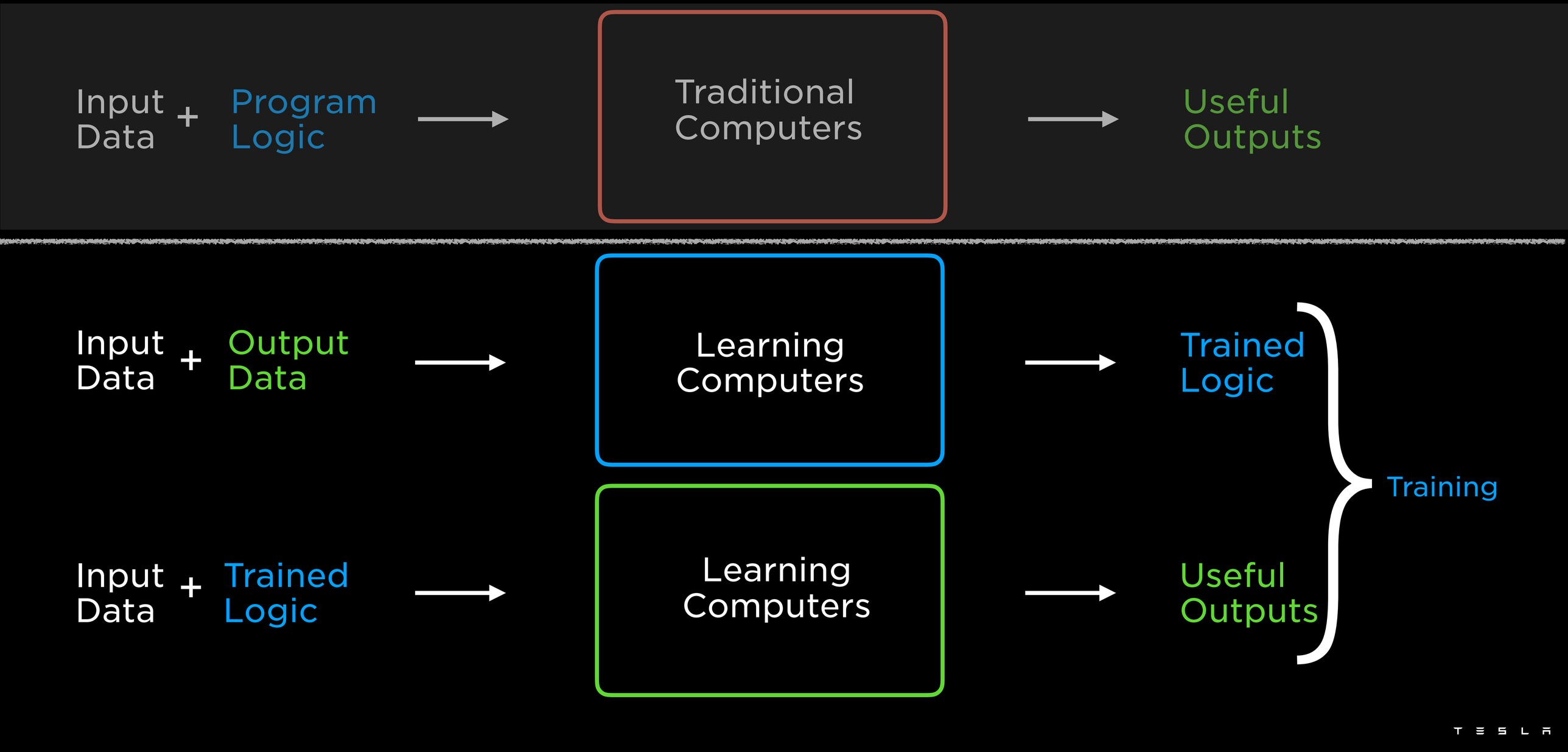
Why Do We Need a Different Compute Platforms?



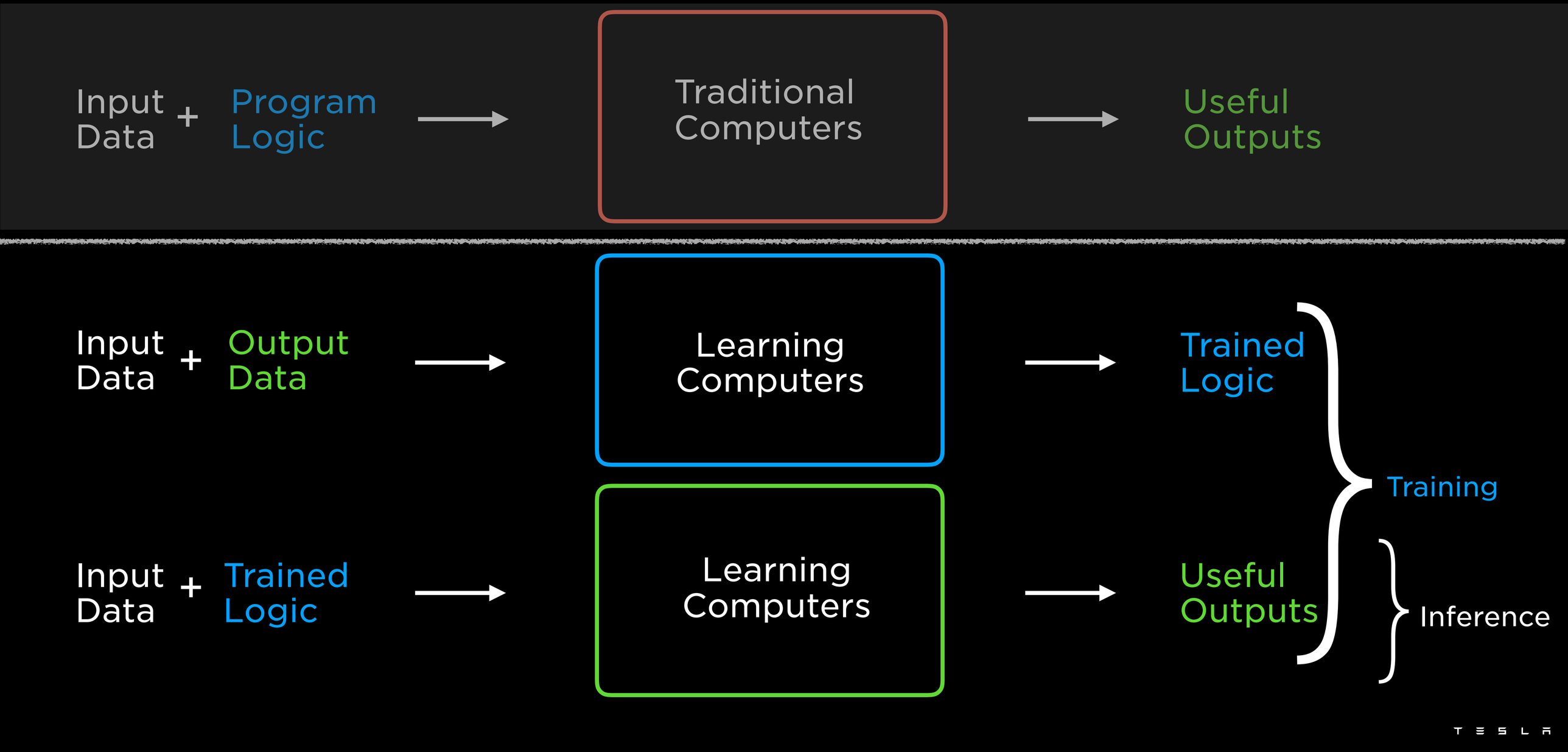
Why Do We Need a Different Compute Platforms?



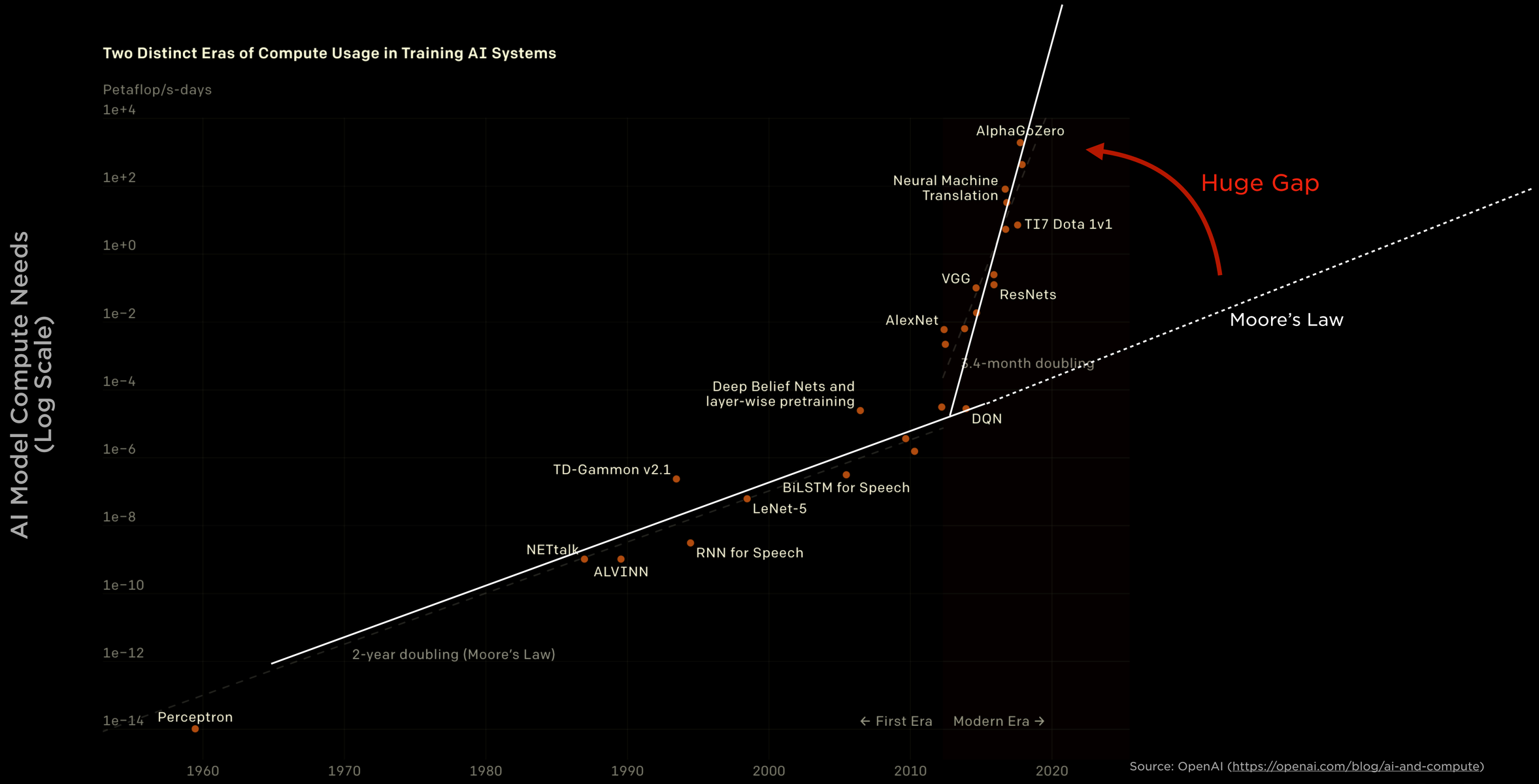
Why Do We Need a Different Compute Platforms?



Why Do We Need a Different Compute Platforms?



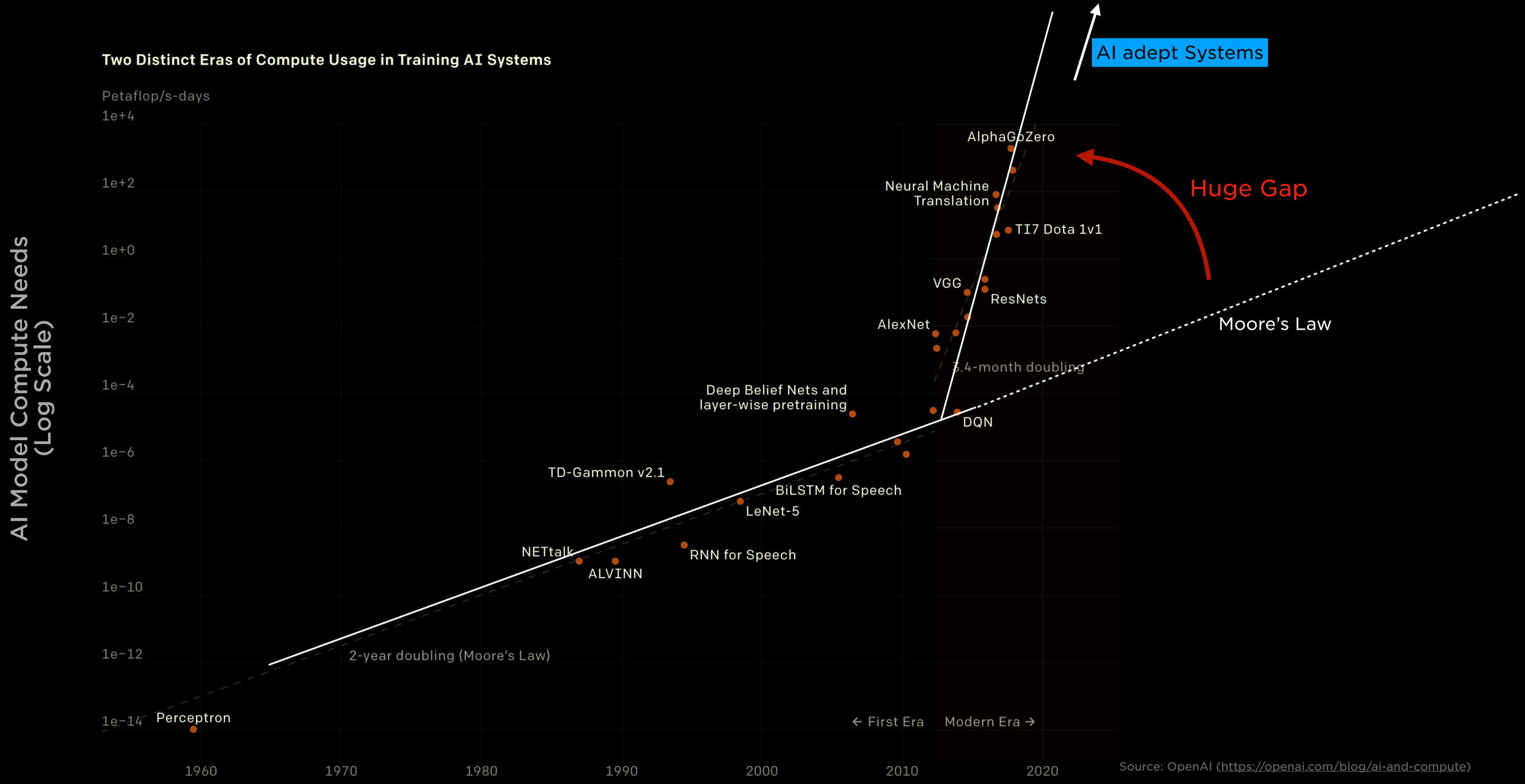
Exponential Rate Gaps in Training Systems



Source: OpenAI (https://openai.com/blog/ai-and-compute)

TESLA

Exponential Rate Gaps in Training Systems



Source: OpenAI (https://openai.com/blog/ai-and-compute)

TESLA

Designing Solutions for AI Level Needs

AI Training Systems

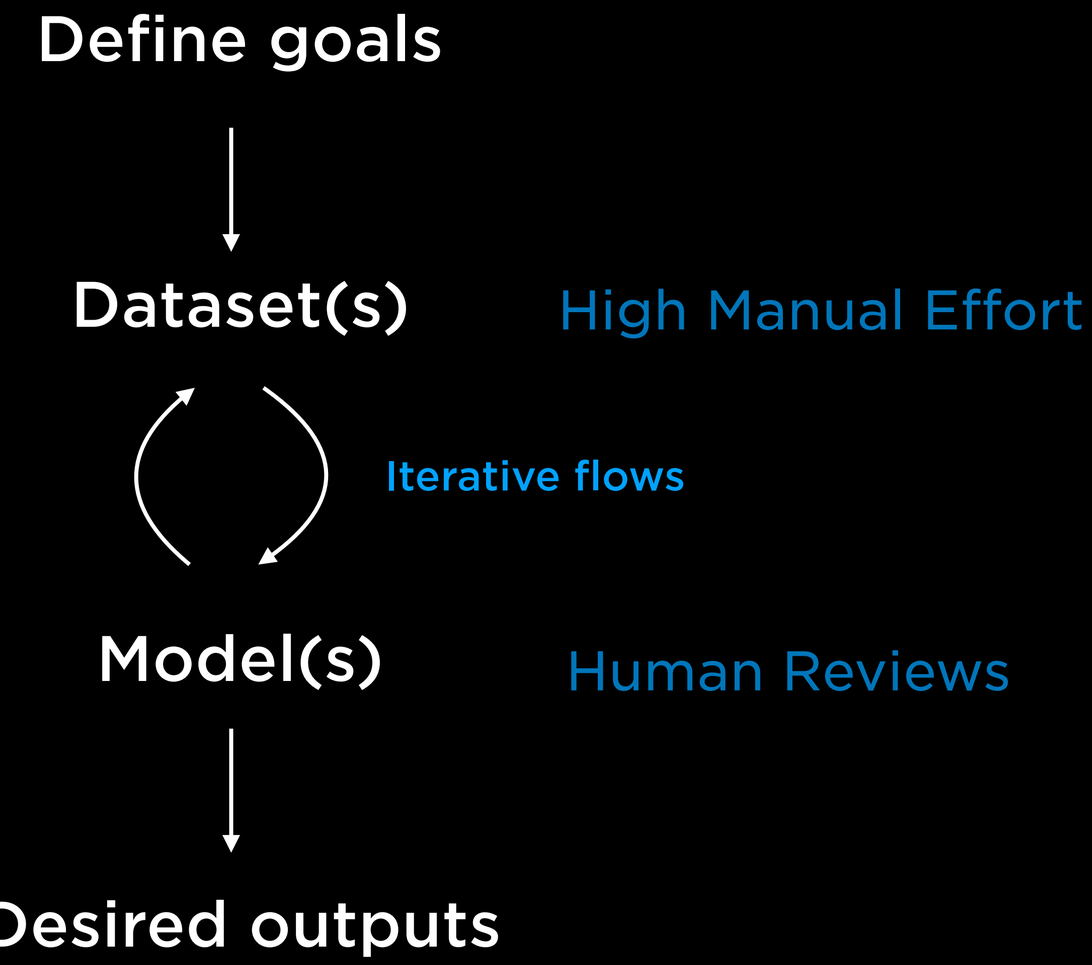
Datasets
Models

SW

Compute
Scale

HW

Typical ML Training Flows



AI Adept Systems

Datasets
Models

SW

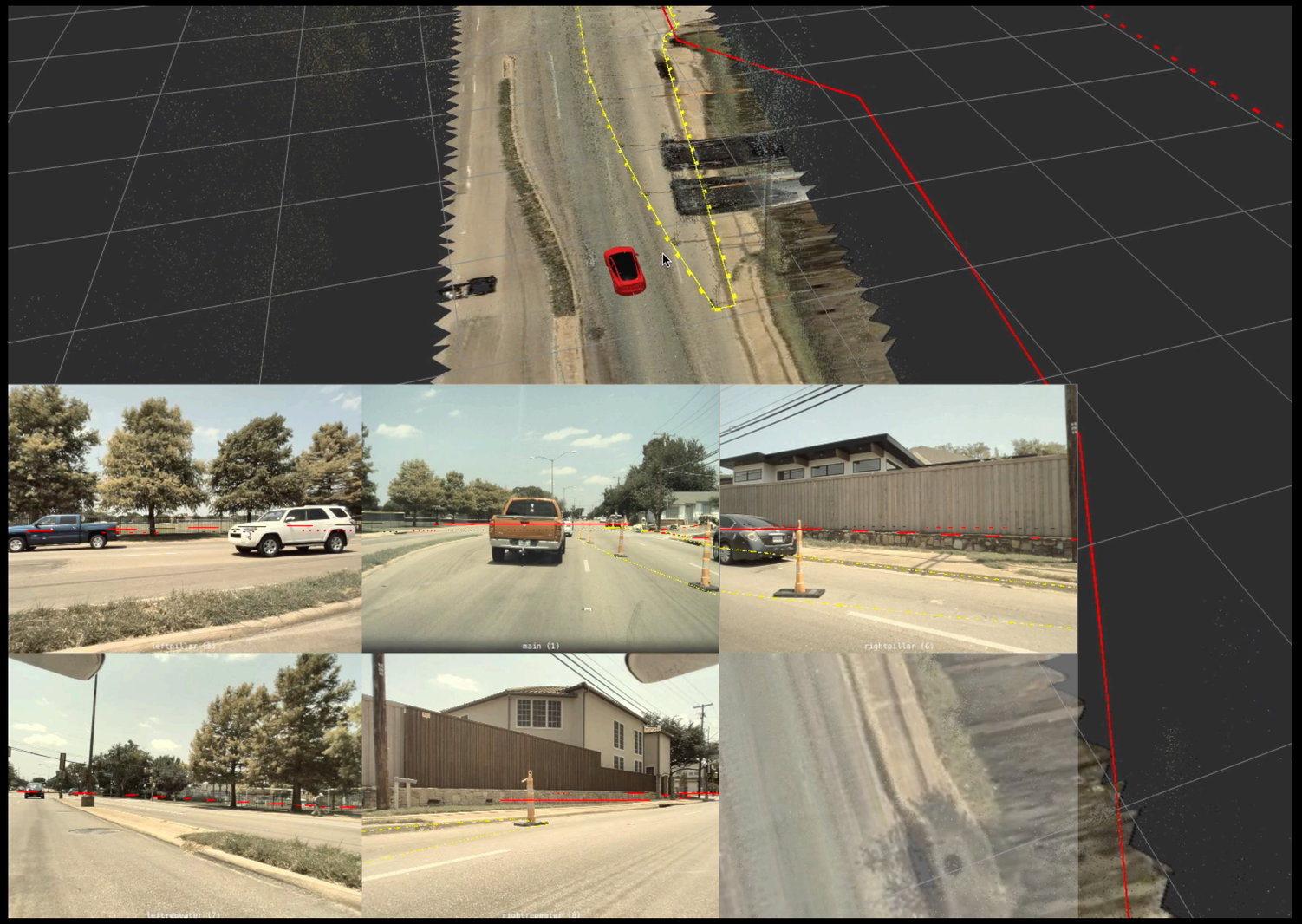
Compute
Scale

HW

2D Image Labeling of Real World Inputs

4D - Space + Time Labeling

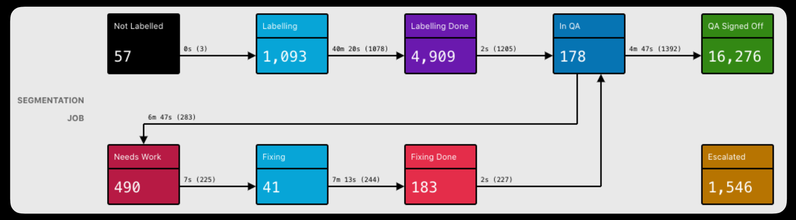
Label Once, Simultaneously Labels All Cameras at Many Frames

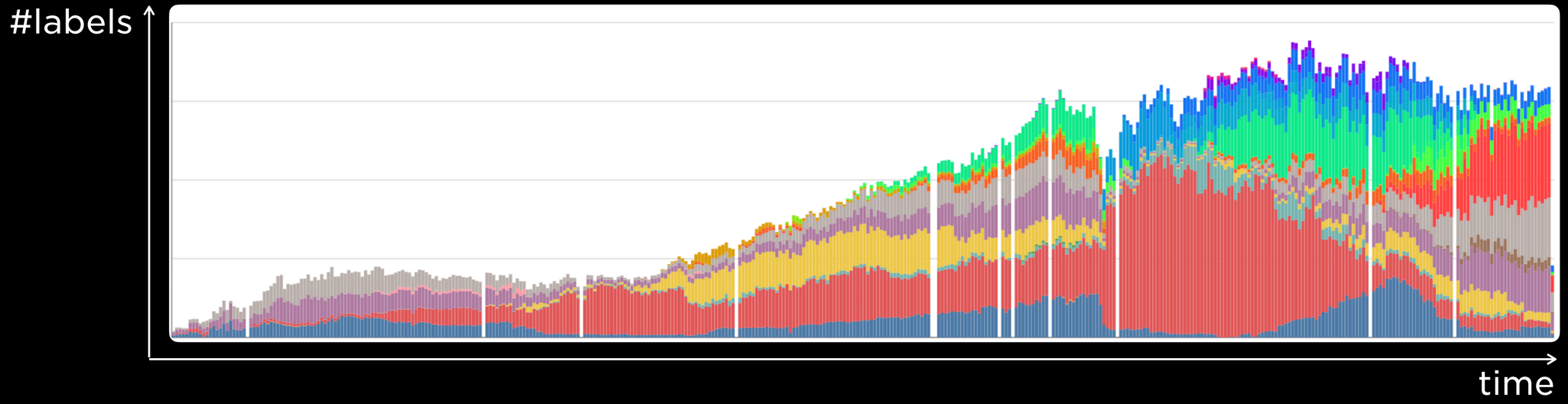


100X Labeling Throughput

Data Labeling Growth

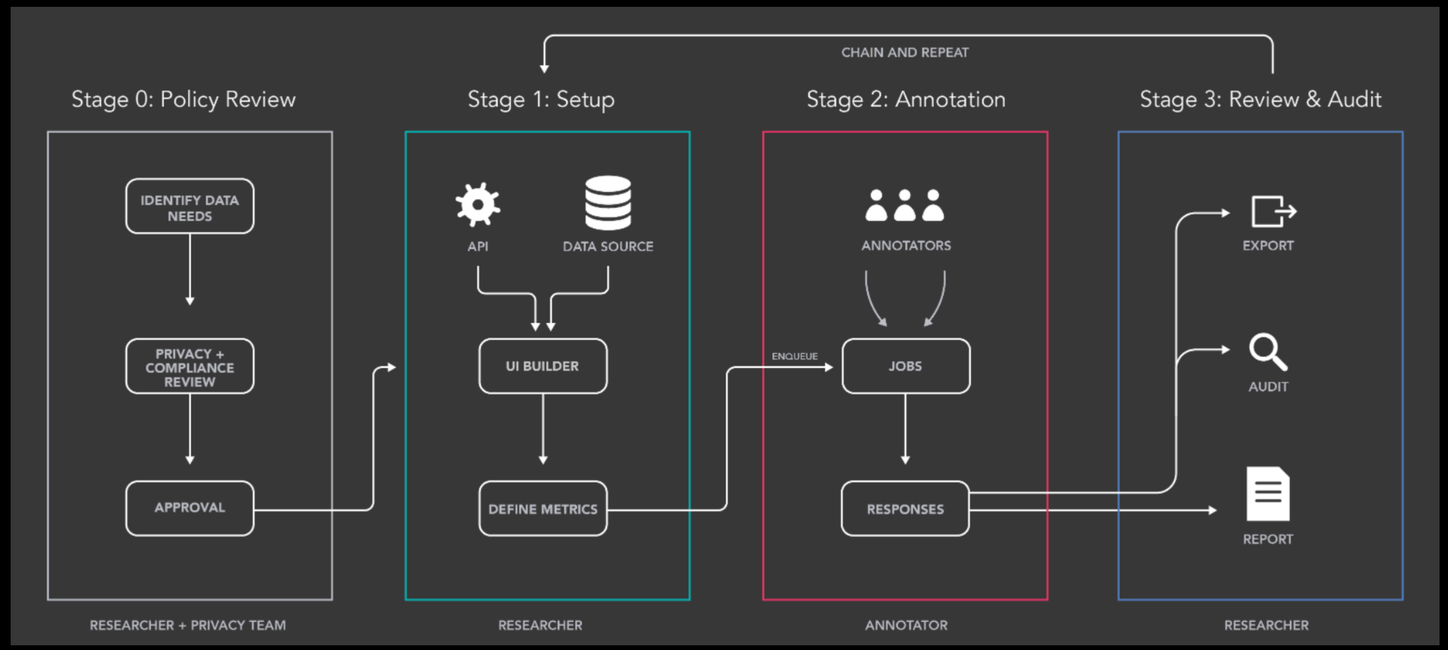
1,000-Person In-House Data Labeling Team
Fully Custom Built Data Labeling & Analytics Infrastructure





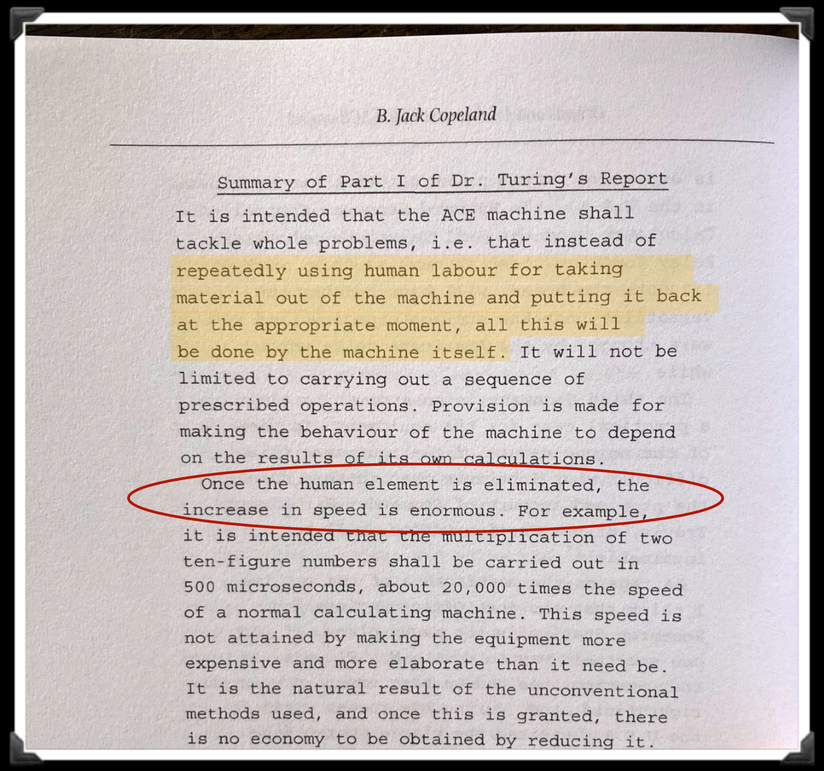
Other Frameworks

"We refer to this framework as **Human-AI loop (Halo.)** researchers can streamline annotation tasks, visualize the results and accuracy metrics of annotations, and export the annotations to start their training modules." : MetaAI



Source: MetaAI - How Facebook Annotates Multimodal Training Data for ML

Historical Parallels



B. Jack Copeland

Summary of Part I of Dr. Turing's Report

It is intended that the ACE machine shall tackle whole problems, i.e. that instead of repeatedly using human labour for taking material out of the machine and putting it back at the appropriate moment, all this will be done by the machine itself. It will not be limited to carrying out a sequence of prescribed operations. Provision is made for making the behaviour of the machine to depend on the results of its own calculations.

Once the human element is eliminated, the increase in speed is enormous. For example, it is intended that the multiplication of two ten-figure numbers shall be carried out in 500 microseconds, about 20,000 times the speed of a normal calculating machine. This speed is not attained by making the equipment more expensive and more elaborate than it need be. It is the natural result of the unconventional methods used, and once this is granted, there is no economy to be obtained by reducing it.

Source:Jack Copeland , Alan Turning's Electronic Brain

"Once the human element is eliminated, the increase in speed is enormous"

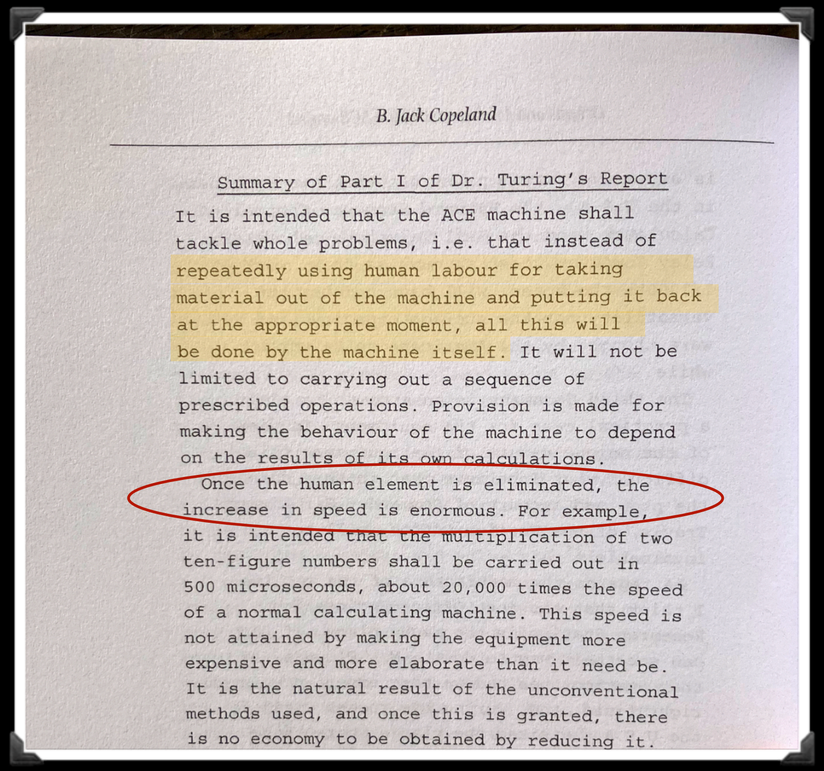
- Alan Turning (As stated in his report from early 1940s)

↓

Led to Advent of programmable computers

Colossus/Eniac/ACE

Historical Parallels



B. Jack Copeland

Summary of Part I of Dr. Turing's Report

It is intended that the ACE machine shall tackle whole problems, i.e. that instead of repeatedly using human labour for taking material out of the machine and putting it back at the appropriate moment, all this will be done by the machine itself. It will not be limited to carrying out a sequence of prescribed operations. Provision is made for making the behaviour of the machine to depend on the results of its own calculations.

Once the human element is eliminated, the increase in speed is enormous. For example, it is intended that the multiplication of two ten-figure numbers shall be carried out in 500 microseconds, about 20,000 times the speed of a normal calculating machine. This speed is not attained by making the equipment more expensive and more elaborate than it need be. It is the natural result of the unconventional methods used, and once this is granted, there is no economy to be obtained by reducing it.

Source:Jack Copeland , Alan Turning's Electronic Brain

"Once the human element is eliminated, the increase in speed is enormous"

- Alan Turning (As stated in his report from early 1940s)

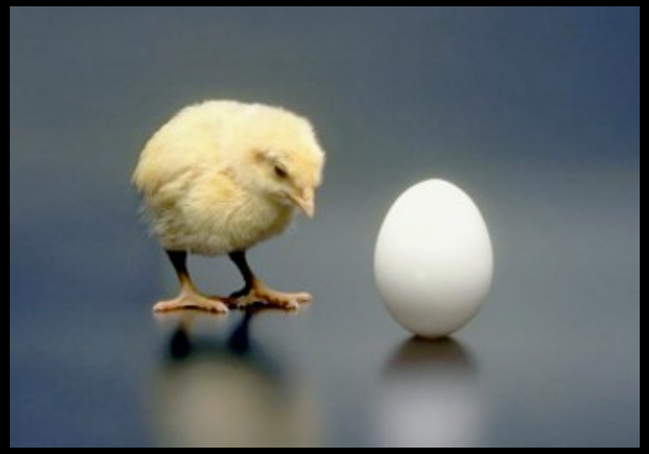
↓

Led to Advent of programmable computers

Colossus/Eniac/ACE

How to do this for AI?

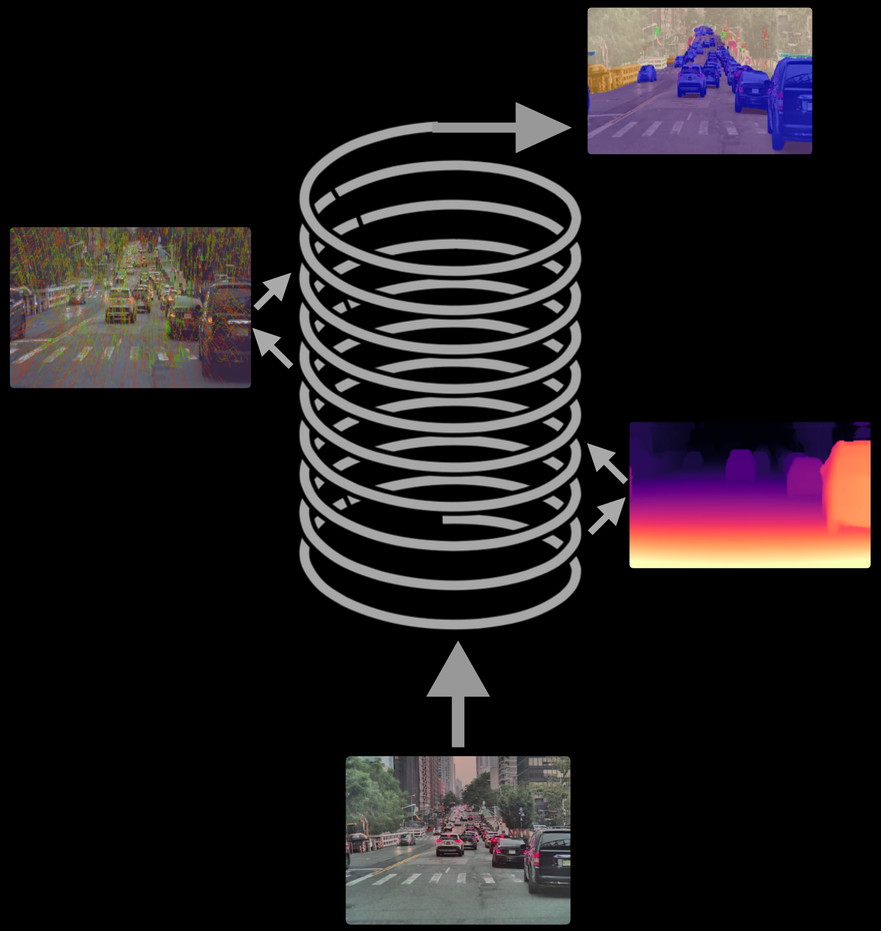
Solution in ML/AI Space Itself !!



Chicken and Egg Problem?

Solution in ML/AI Space Itself !!

Recursive Loops

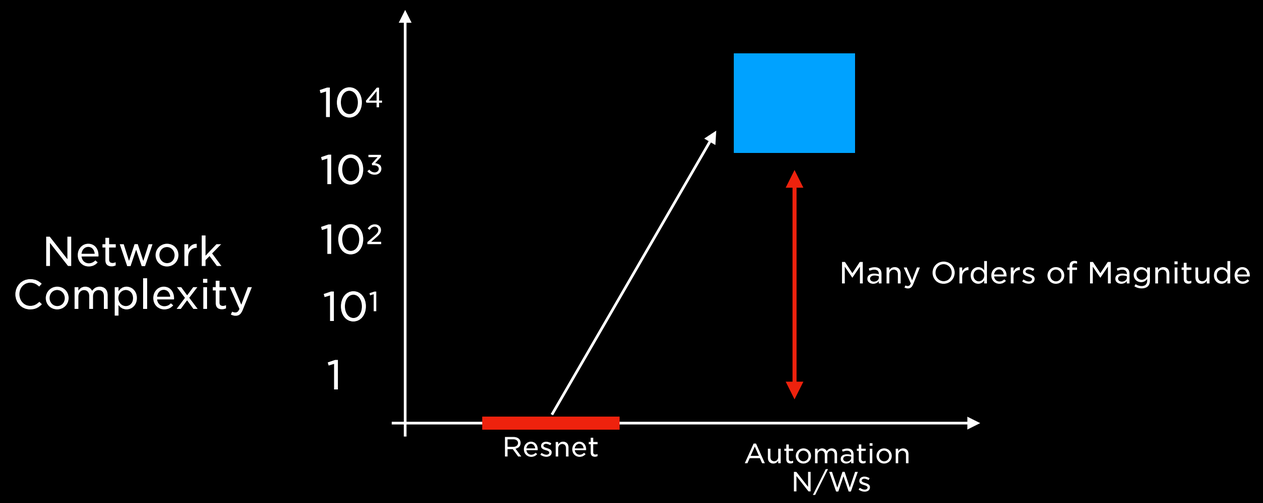


Solution in ML/AI Space Itself !!

Dataset Labeling Automation

Use of Offline Models for Real World Dataset Curation/Labeling

With Reduced Human Loop Dependence

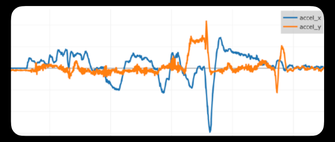


Autolabeling

Clip

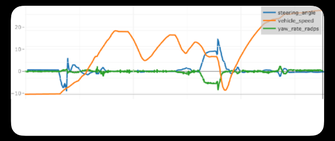
Videos

IMU



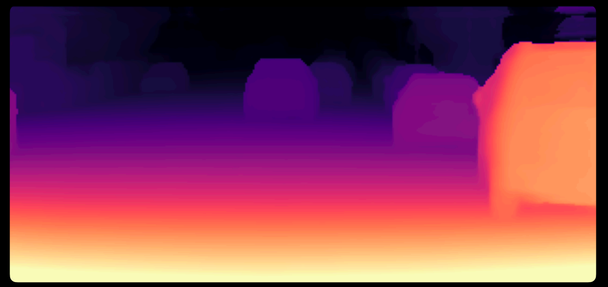
GPS

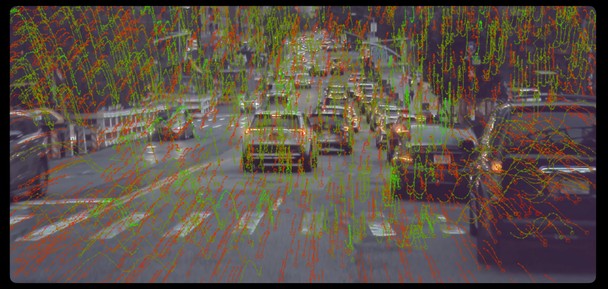
Odometry



...

Offline Neural Networks



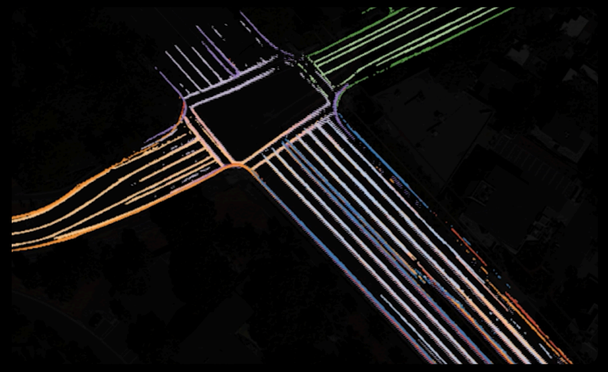


...

Ego Trajectory & Static World Reconstruction

Moving Objects & Kinematics

Labels



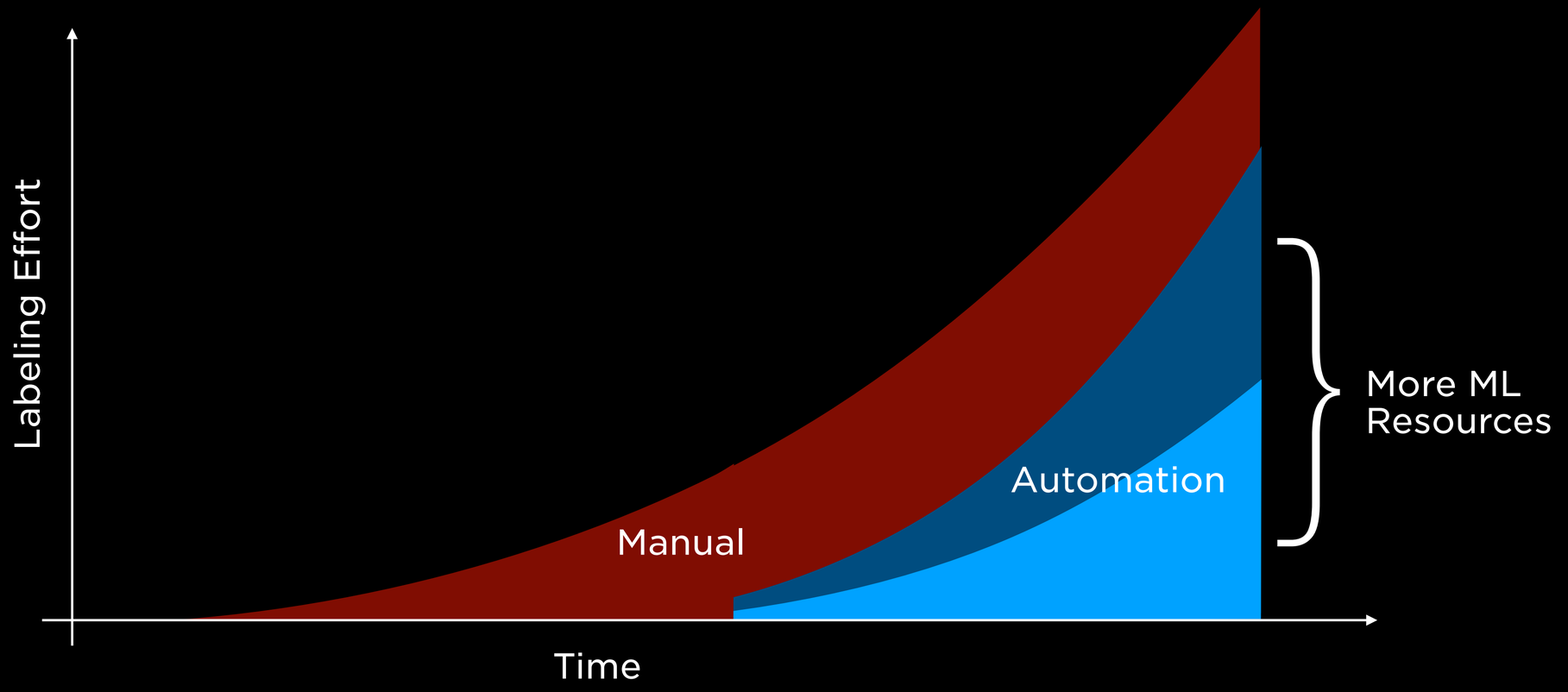
Ask the Test Fleet for Interesting Clips

```
"name": "cipv-low-vis",
"requester": "img-vid-cipv-low-vis-{{seq}}",
"description": "Low visibility with a CIPV",
"query": {
   "$and": [
      {"$eq": [
         {"$decimate":
            {"$conv":
               {"$and": [
                  {"$eq": ["{{active-gear}}", 4]}, // In drive
                  {"$not": "@VisionSceneTags.main.scene_tag_array[13].activated"}, // GARAGE_DOOR_CLOSED
                  {"$not": "@VisionSceneTags.main.scene_tag_array[15].activated"}, // INDOOR
                  {"$gt": "@TelemetryOutput.distance_travelled_m", 1000},
                  {"$not": "@lss_app.right_lane.lane_change"}, // No right lane change
                  {"$not": "@lss_app.left_lane.lane_change"}, // No left lane change
                  {"$not": "@moving_object_output[0].cutin_active_in_scene"}, // No cutin
                  {"$lt": "@moving_object_output[0].max_region_tag_cutin_prob", 0.1},
                  {"$lt": "@moving_object_output[2].max_region_tag_cutin_prob", 0.1},
                  {"$gt": "{{veh-speed-mps}}", 2.2} // 5 mph
               ]},
               "h": [1,1,1,1,1, 1,1,1,1,1]  // 10s
            },
            "N": 50, // 1s period
            "stateless-child": true
         },
         10
      ]},
```

And the Test Fleet Giveth Back

10k Such Clips Collected & Automatically Labelled Within a Week

Investments for Offline Dataset Speedup



AI Adept Systems

Datasets
Models

SW

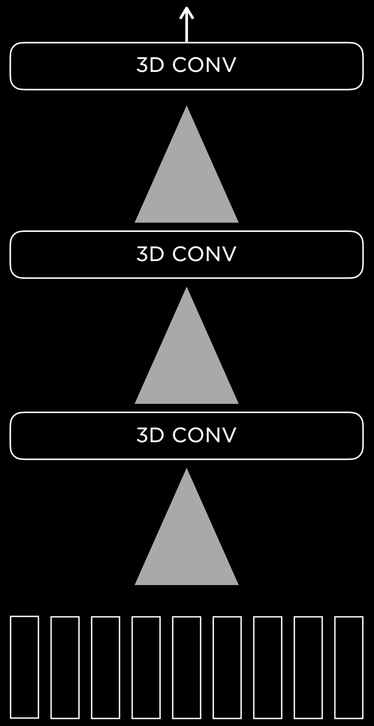
Compute
Scale

HW

Video Training Modules

3D CONV

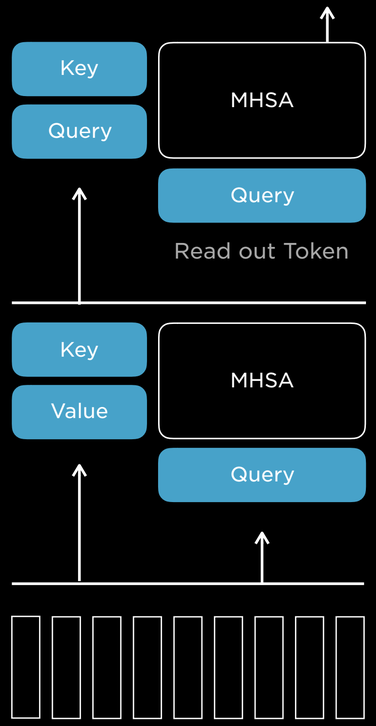
20x80x300
W x H x C



20x80x300x12
W x H x C x T

Transformer

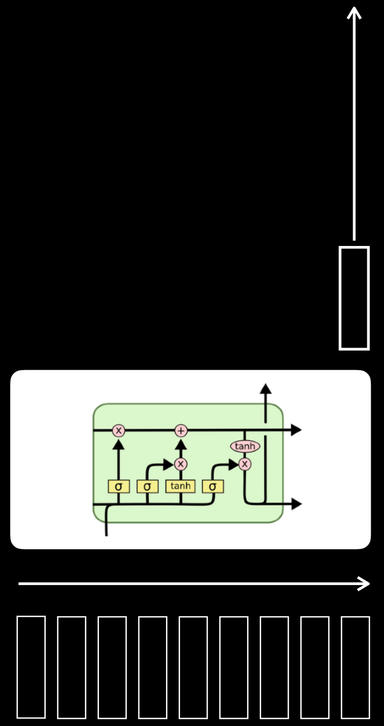
20x80x300
W x H x C



20x80x300x12
W x H x C x T

Recurrent Neural Net

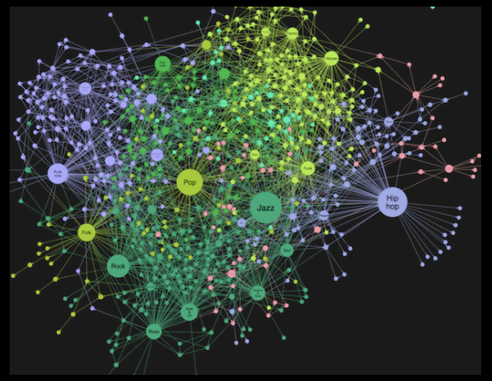
20x80x300
W x H x C



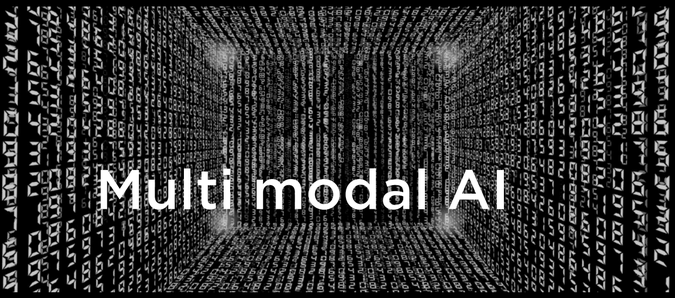
20x80x300x12
W x H x C x T

Many More in Research

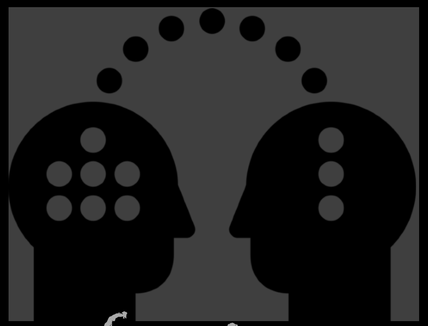
KNOWLEDGE GRAPHS



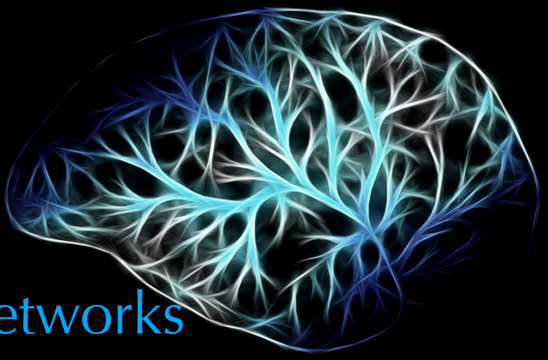
BigGAN



Semantic networks



Transfer Learning



ART Networks

Image Source: Google Images

AI Adept Systems

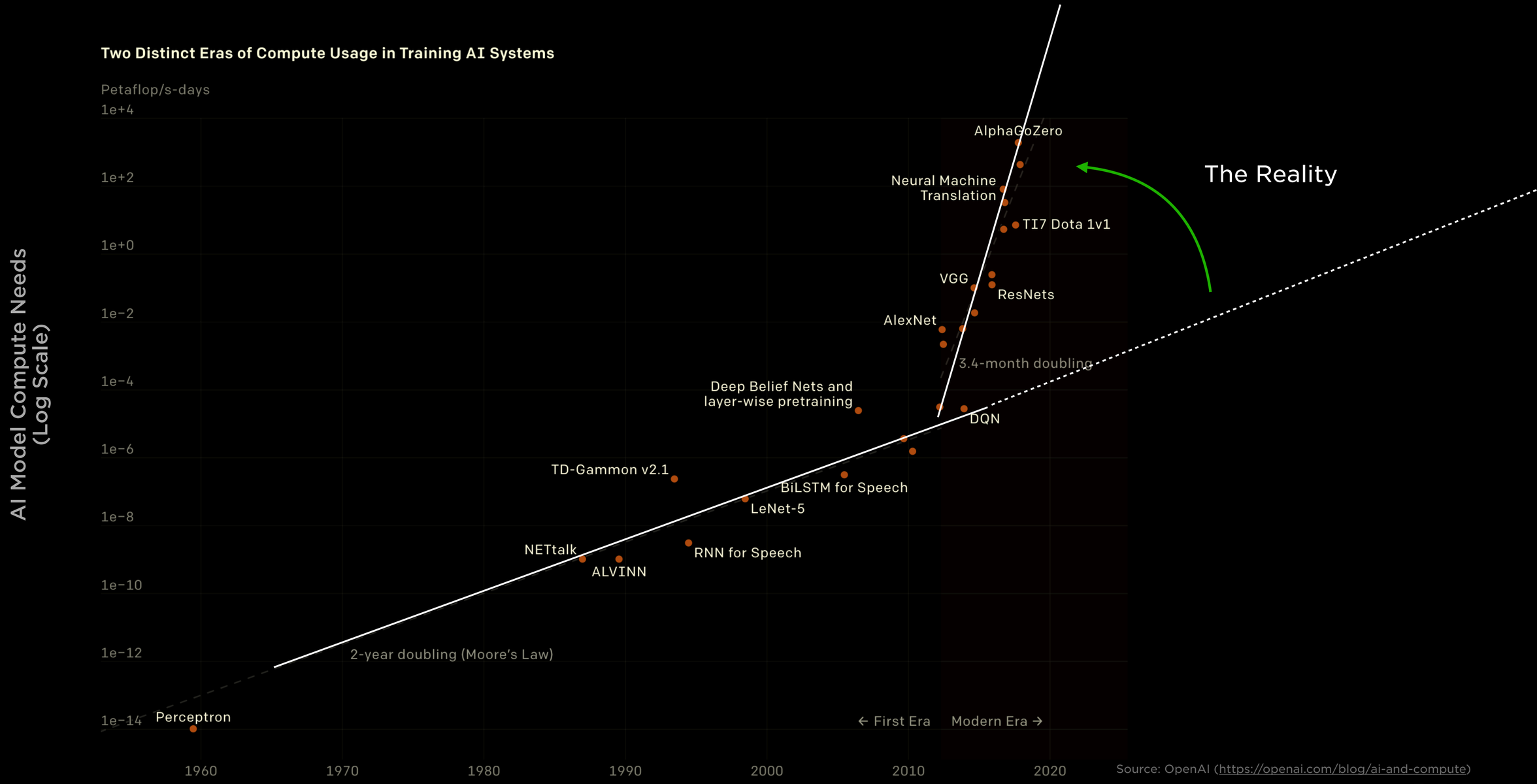
Datasets
Models

SW

Compute
Scale

HW

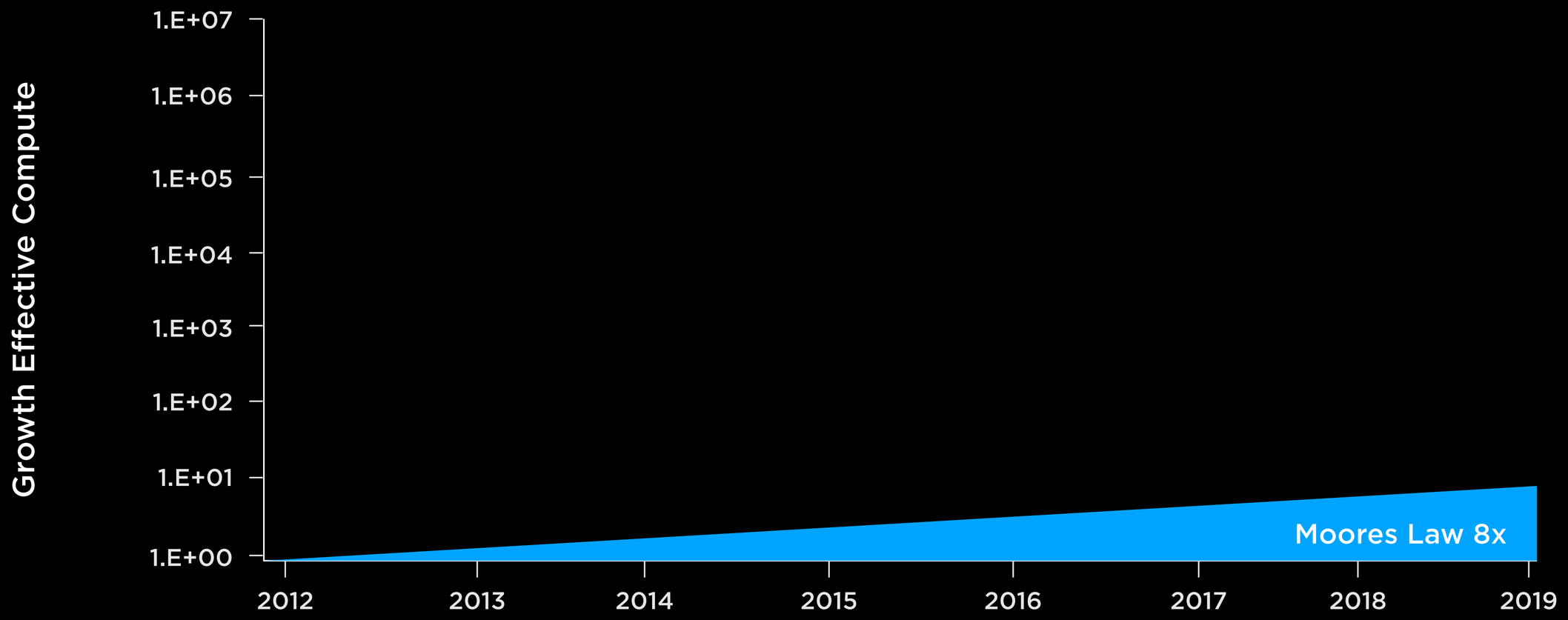
Exponential Rate Gaps in Training Systems



Source: OpenAI (https://openai.com/blog/ai-and-compute)

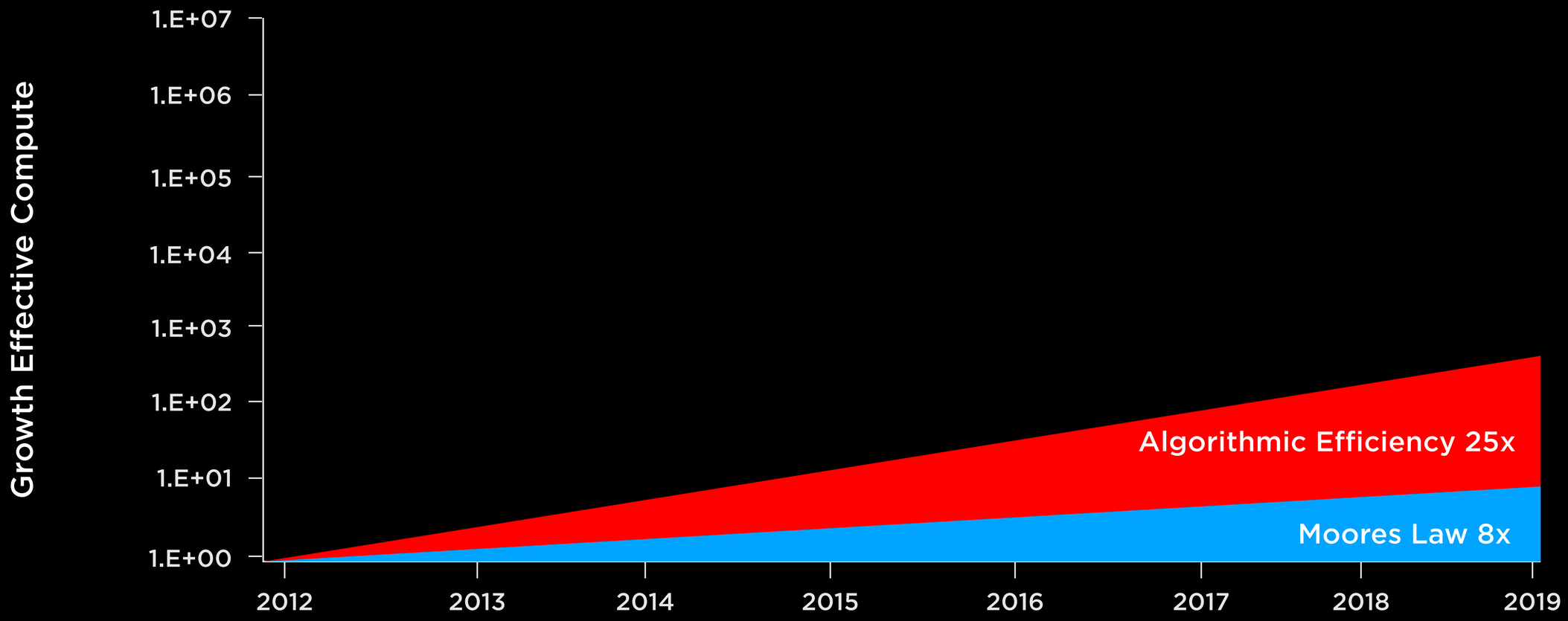
TESLA

Effective Compute Over Time



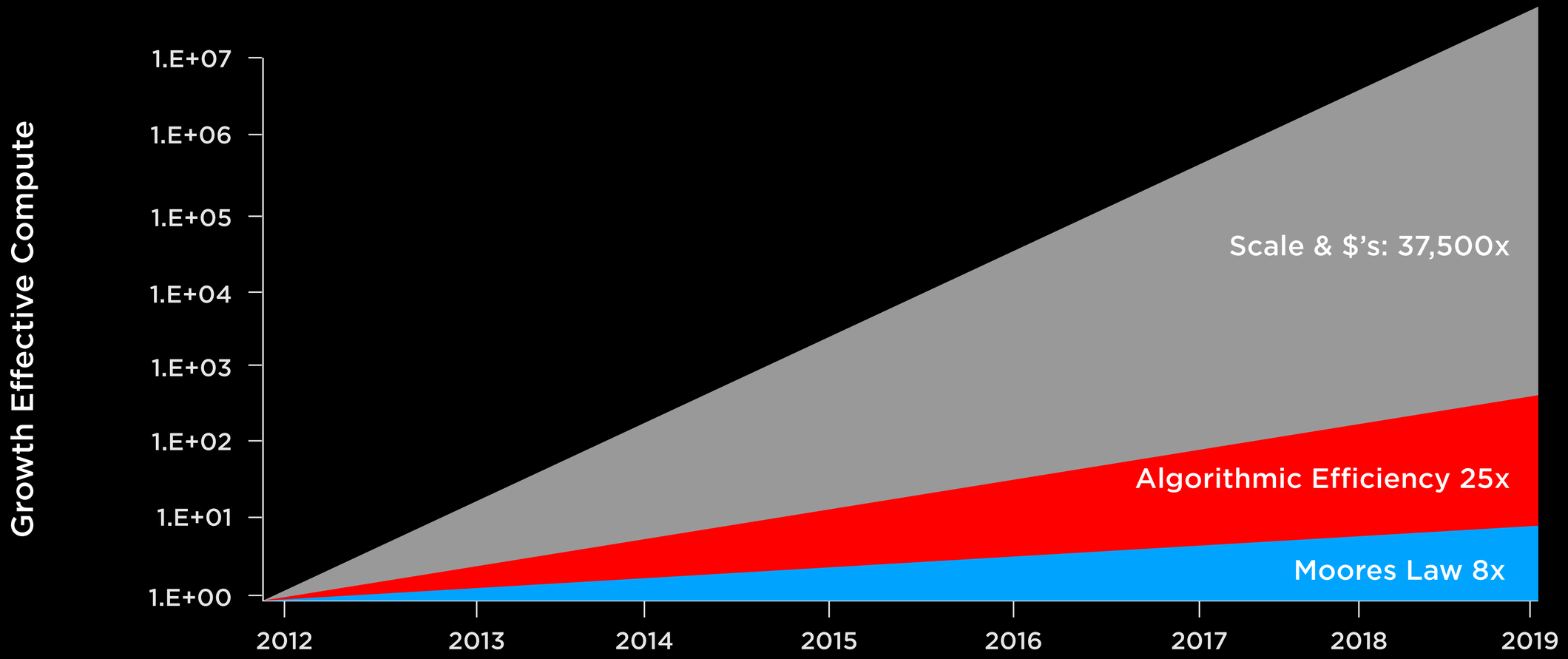
Source : D Hernandez, T Brown, OpenAI : Measuring the Algorithmic Efficiency of Neural Networks

Effective Compute Over Time



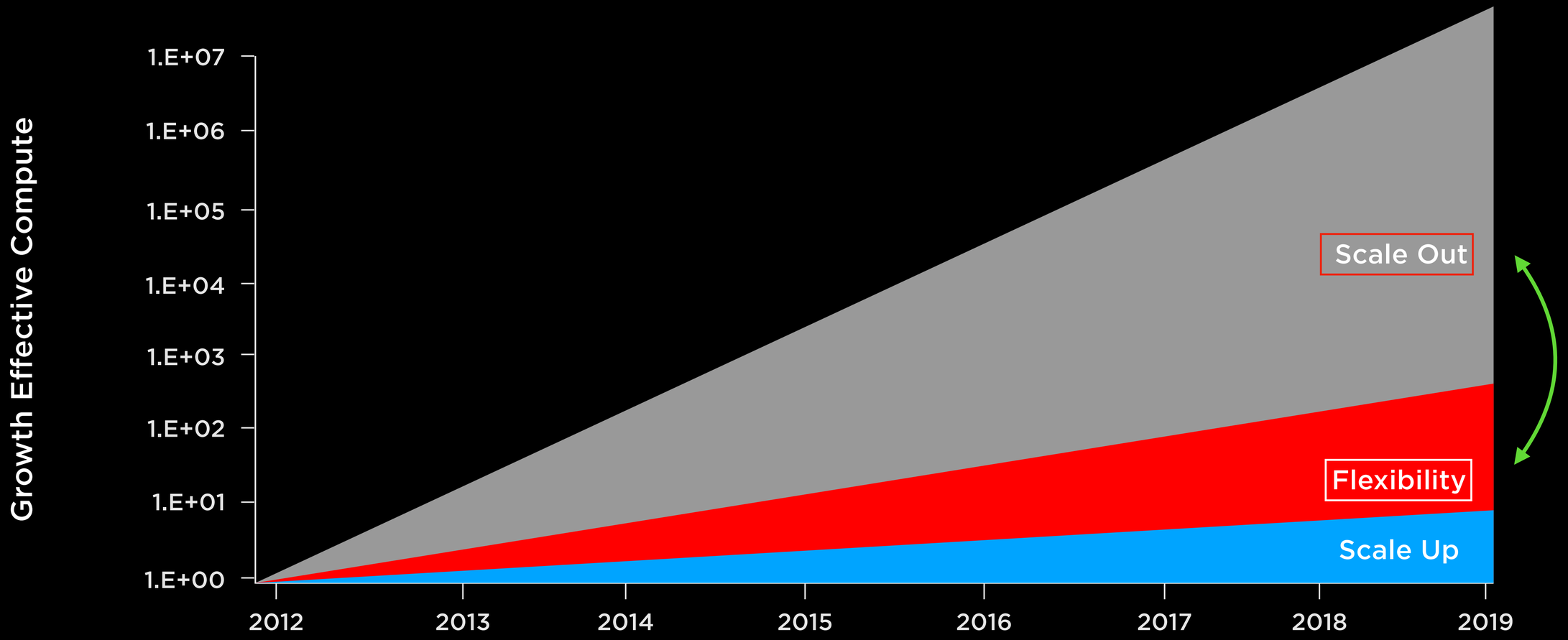
Source : D Hernandez, T Brown, OpenAI : Measuring the Algorithmic Efficiency of Neural Networks

Effective Compute Over Time



Source : D Hernandez, T Brown, OpenAI : Measuring the Algorithmic Efficiency of Neural Networks

Effective Compute Over Time



Source : D Hernandez, T Brown, OpenAI : Measuring the Algorithmic Efficiency of Neural Networks

Climbing Up to ML

Big Data

2005-2010

Climbing Up to ML

Big Compute 2010-2018

Big Data

Climbing Up to ML

Big Models

Big Compute

Big Data

2018-Now

Machine Learning

Climbing Up to AI

Realworld Data

Big Models

Big Compute

Big Data

Climbing Up to AI

Better Compute

Realworld Data

Big Models

Big Compute

Big Data

Climbing Up to AI

...

...

Better Models

Better Compute

Realworld Data

Big Models

Big Compute

Big Data

AI System Traits

Flexible Compute

Real world datasets

Gigantic models

Huge Scale out

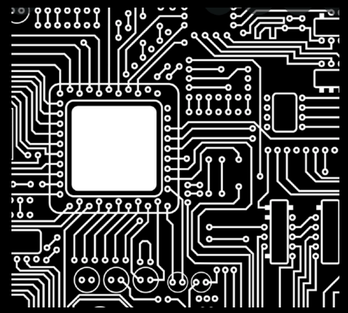
Real time performance

Feed the beast

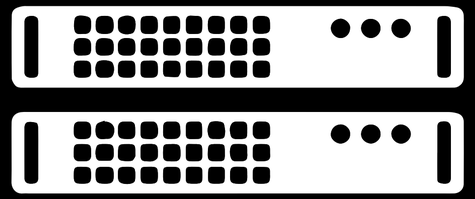
Traditional Hierarchies

Chip

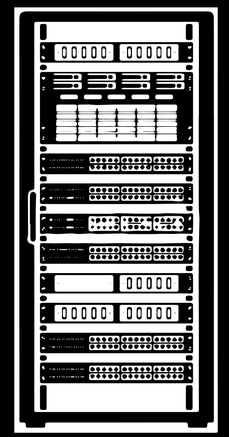
Package



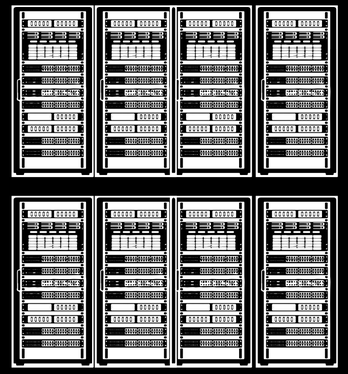
Boards



Boxes

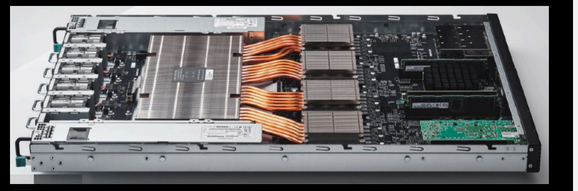


Racks

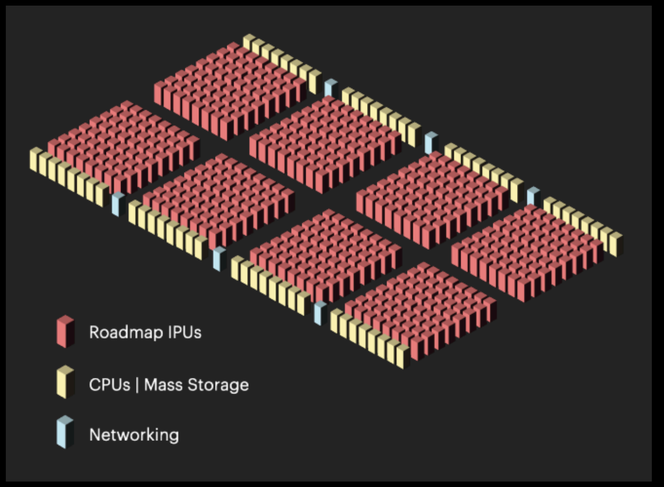


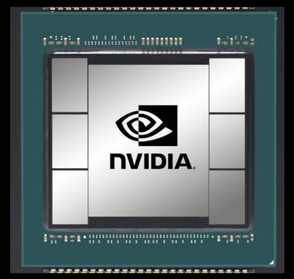
Datacenter/
Buildings

Example Hierarchy



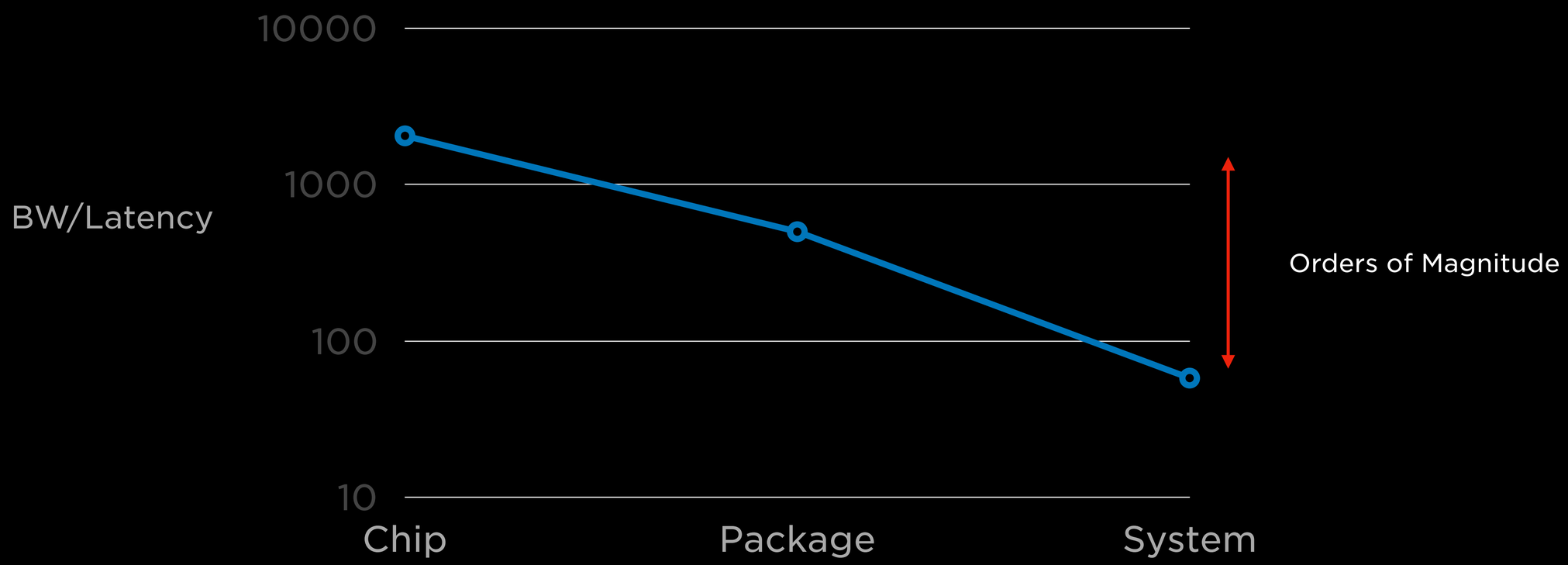
Source: Graphcore



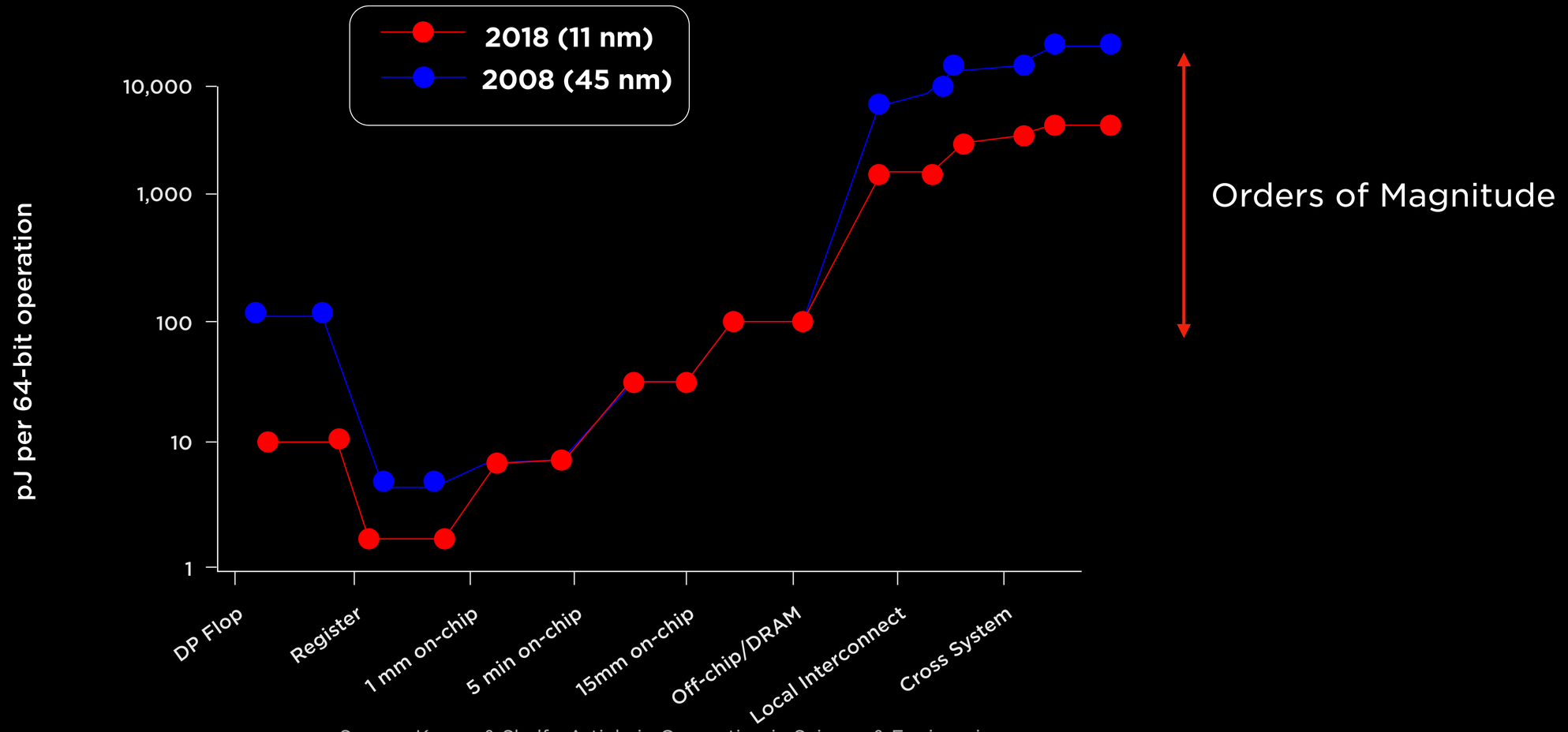


Source: Nvidia

Traditional BW & Latency Scaling Discontinuities



Traditional Hierarchy Power

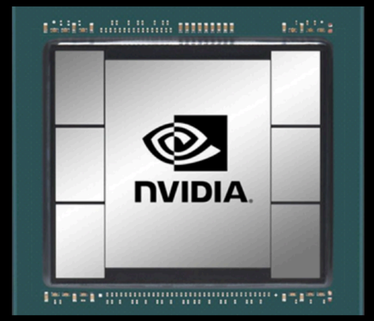


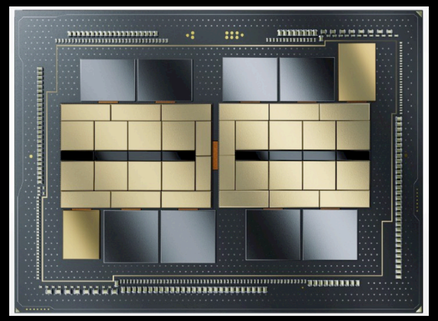
Source: Kogge & Shalf - Article in Computing in Science & Engineering

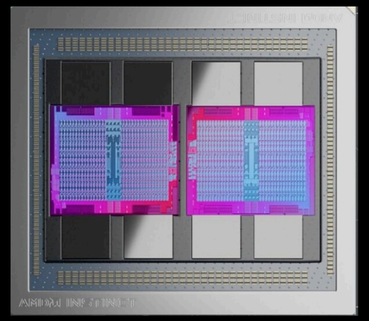
Mitigation for Integration Hierarchy Discontinuities

Reticle Sized Dies

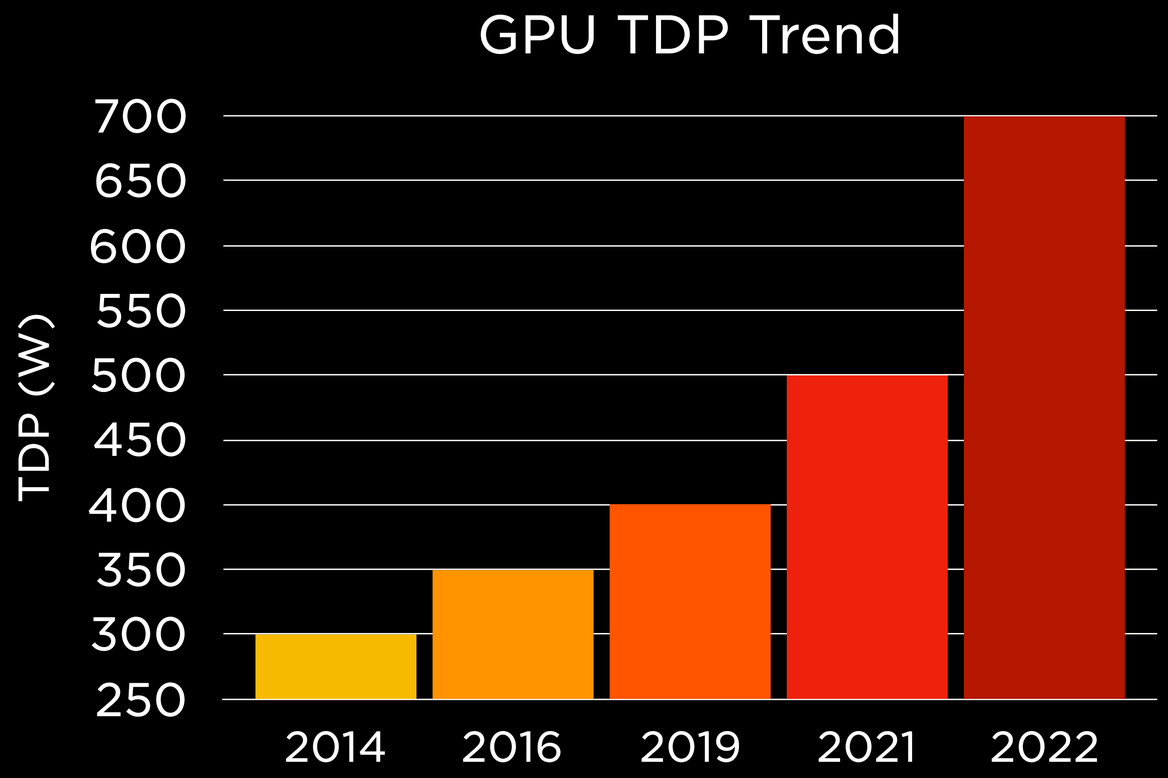
2-3x Reticle Sized Interposers/EMIBs/MCMs



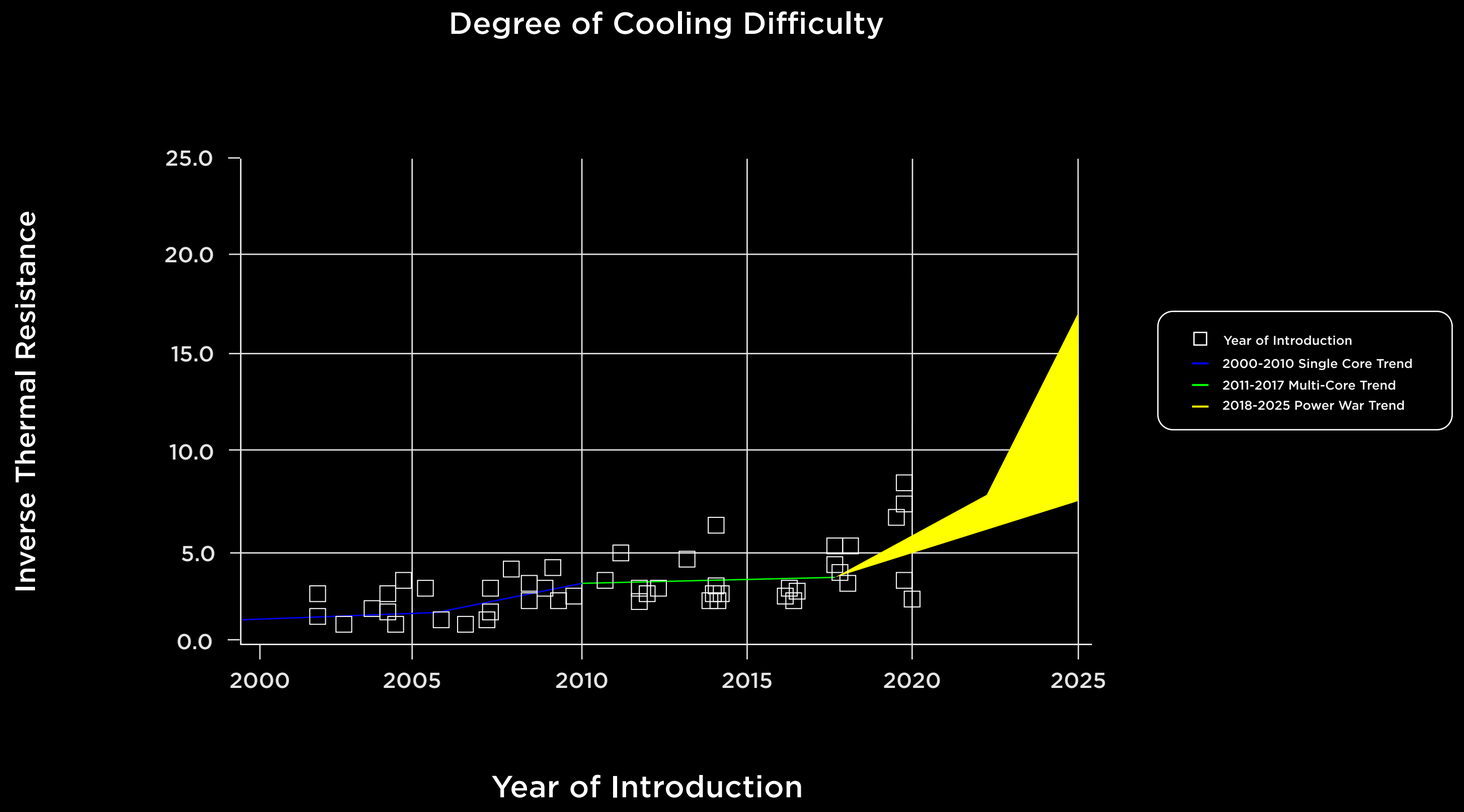




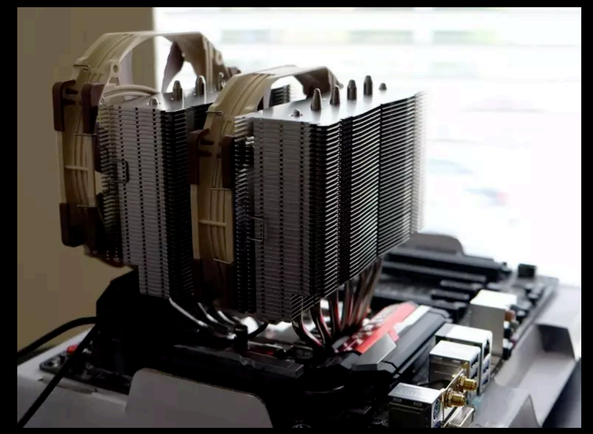
Power Trend


GPU TDP Trend
TDP (W)
700
650
600
550
500
450
400
350
300
250
2014
2016
2019
2021
2022

Cooling Difficulties



Source: White-paper on Emergence and Expansion of Liquid Cooling in Mainstream Data Centers



https://www.shacknews.com/

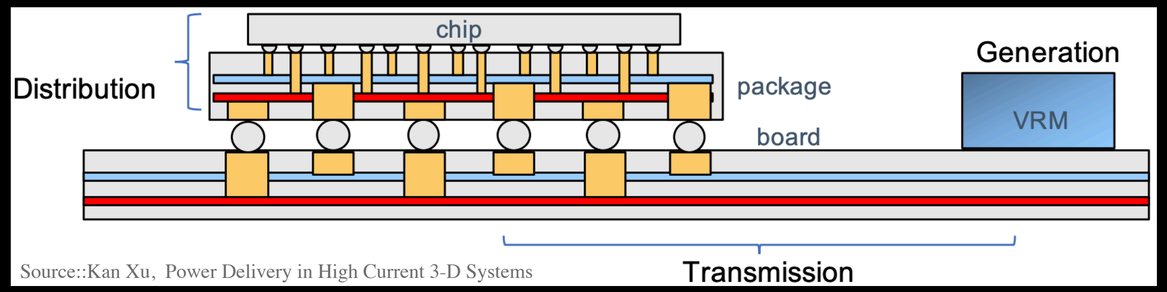
2016

3U (~135mm)

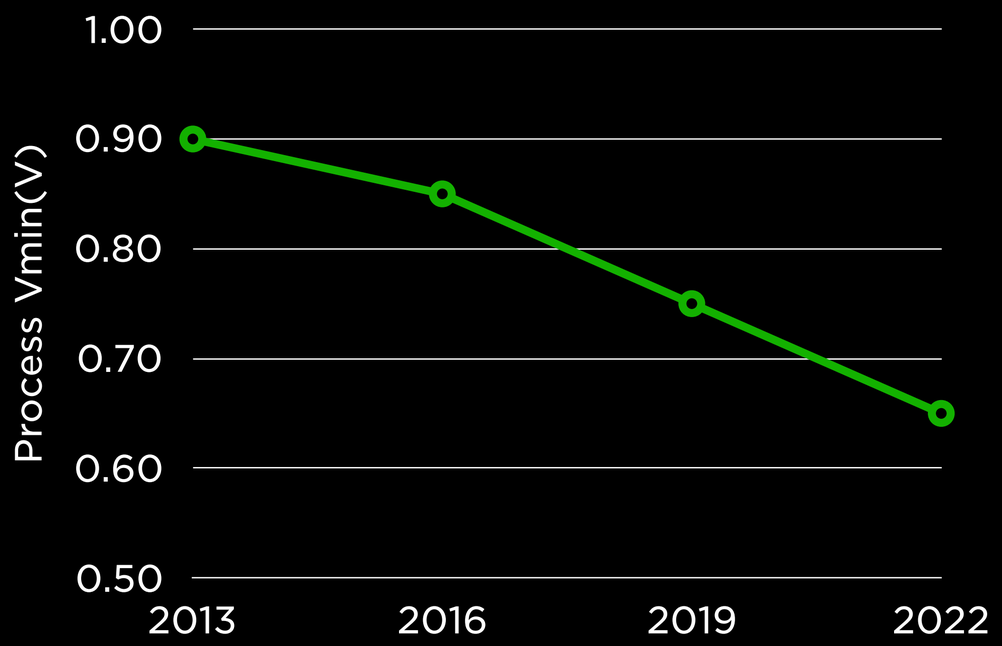
2022

8U (~356mm)

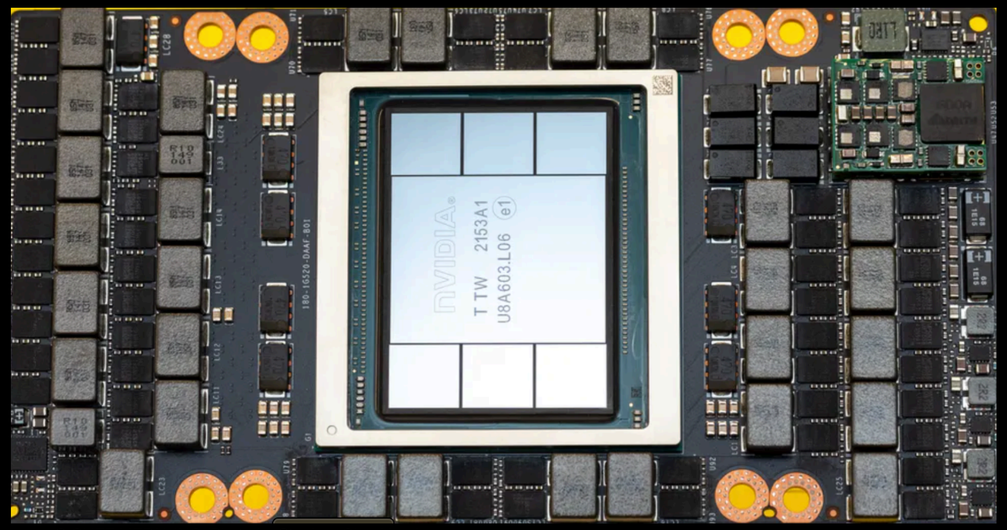
Lateral Power Delivery Challenges



Source::Kan Xu, Power Delivery in High Current 3-D Systems

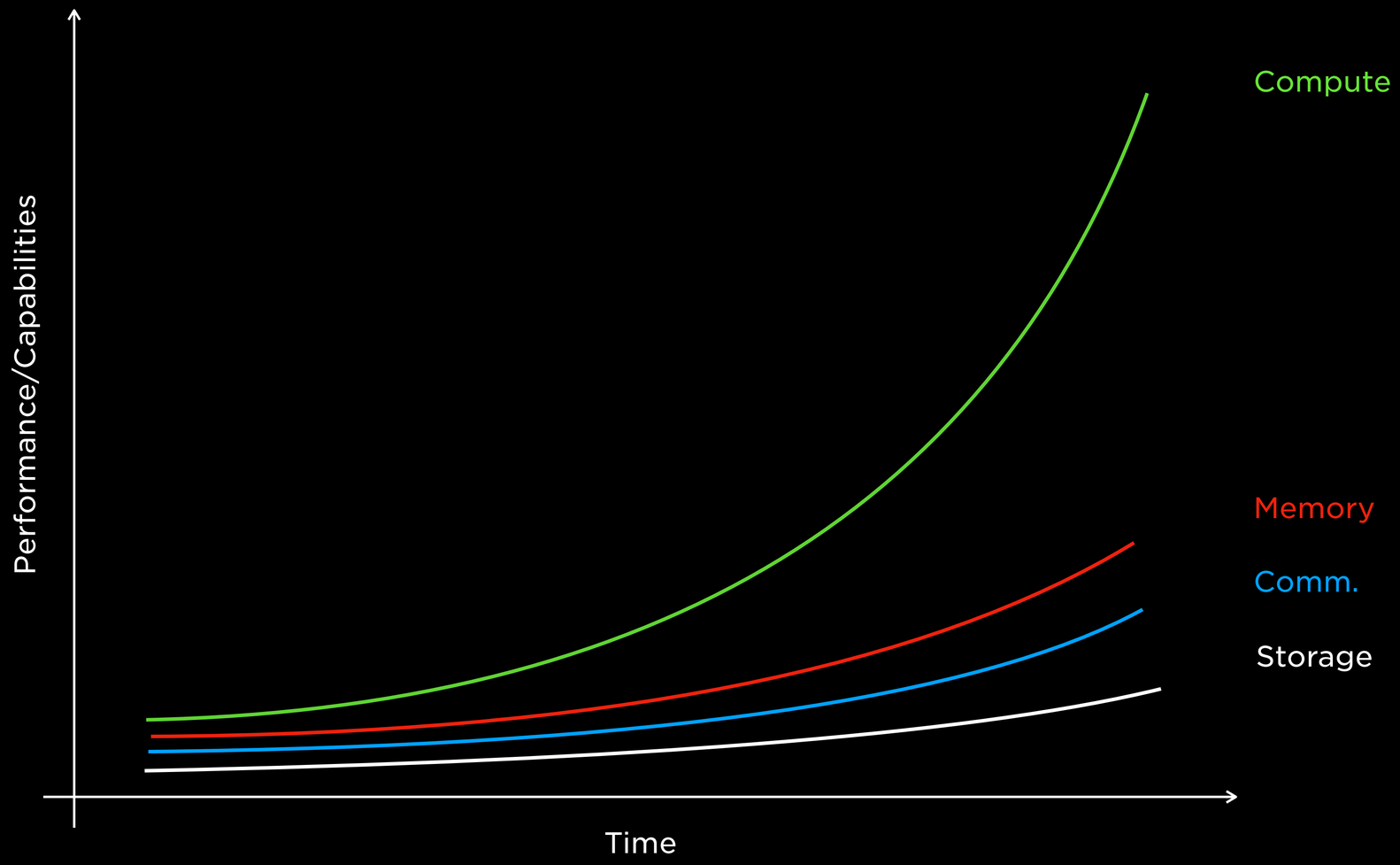


Source: www.nextplatform.com

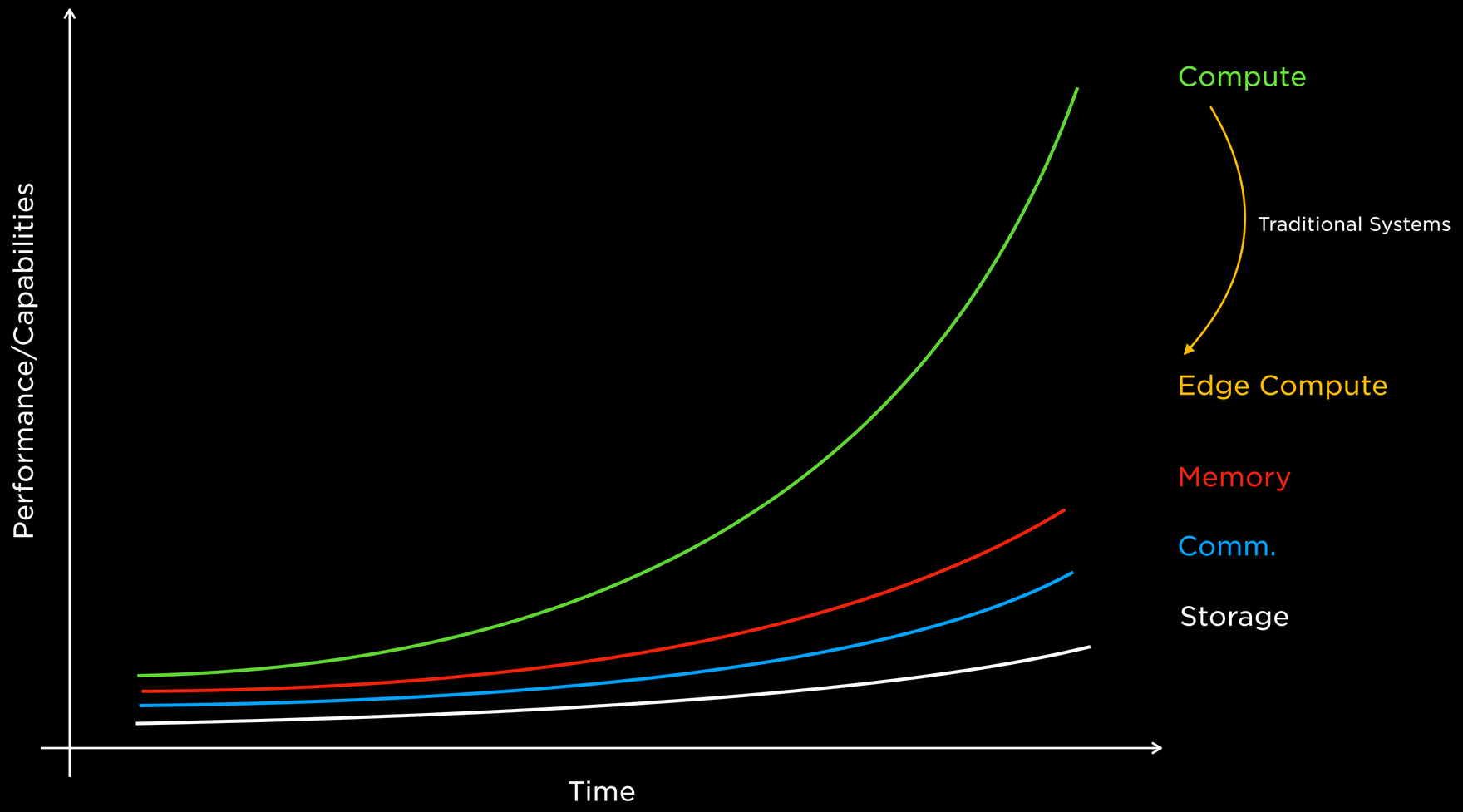


© CNET MEDIA INC./Photo by Stephen Shankland

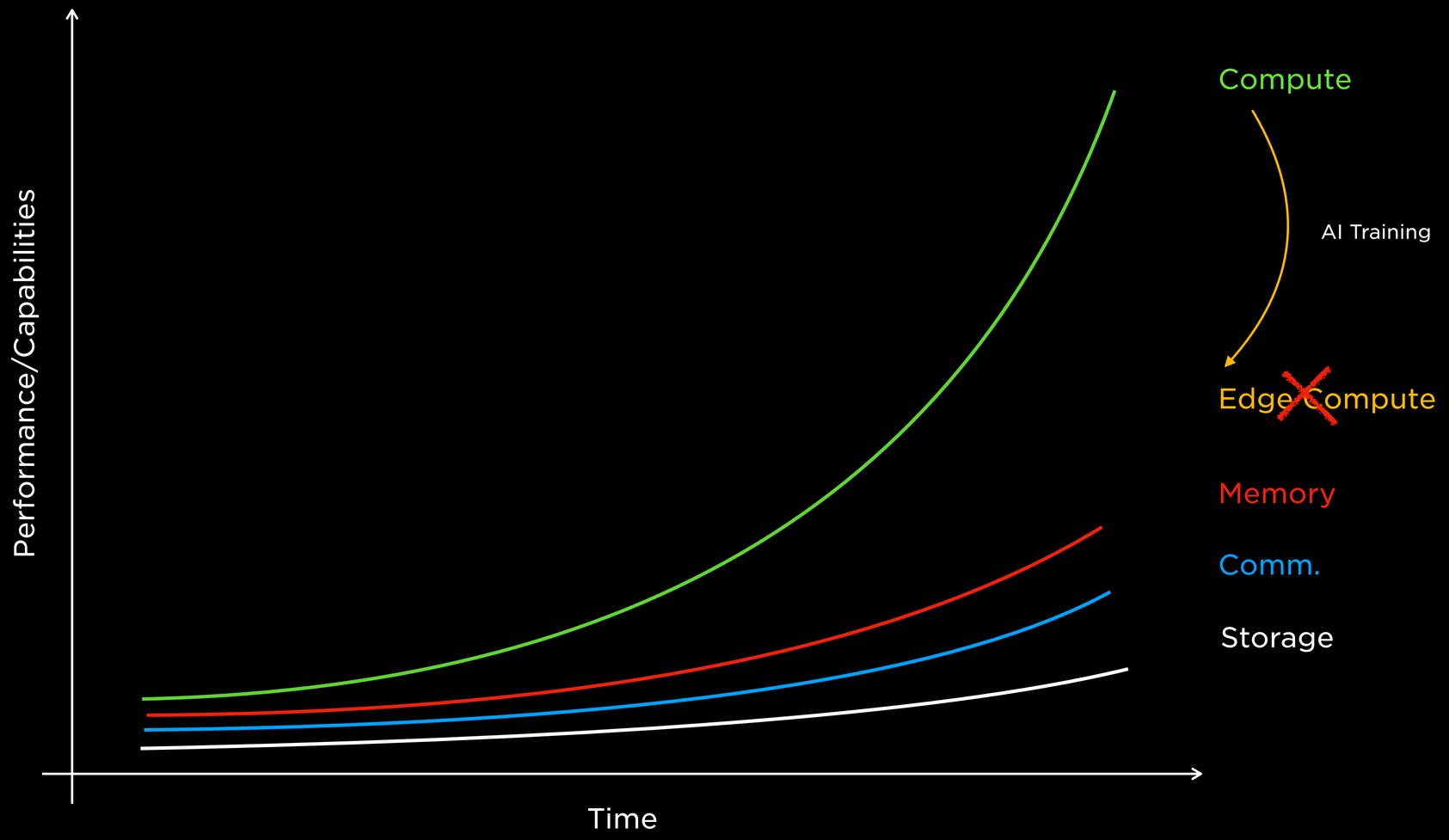
Tech Scaling Gaps



Tech Scaling Gaps



Tech Scaling Gaps



Scaling Gaps for Efficient Scaleout

Discrete/Chip Centric Approaches Are Severely Limiting Value Propositions

BW & Latency Losses

Power To Traverse Hierarchies

Device vs I/O Scaling

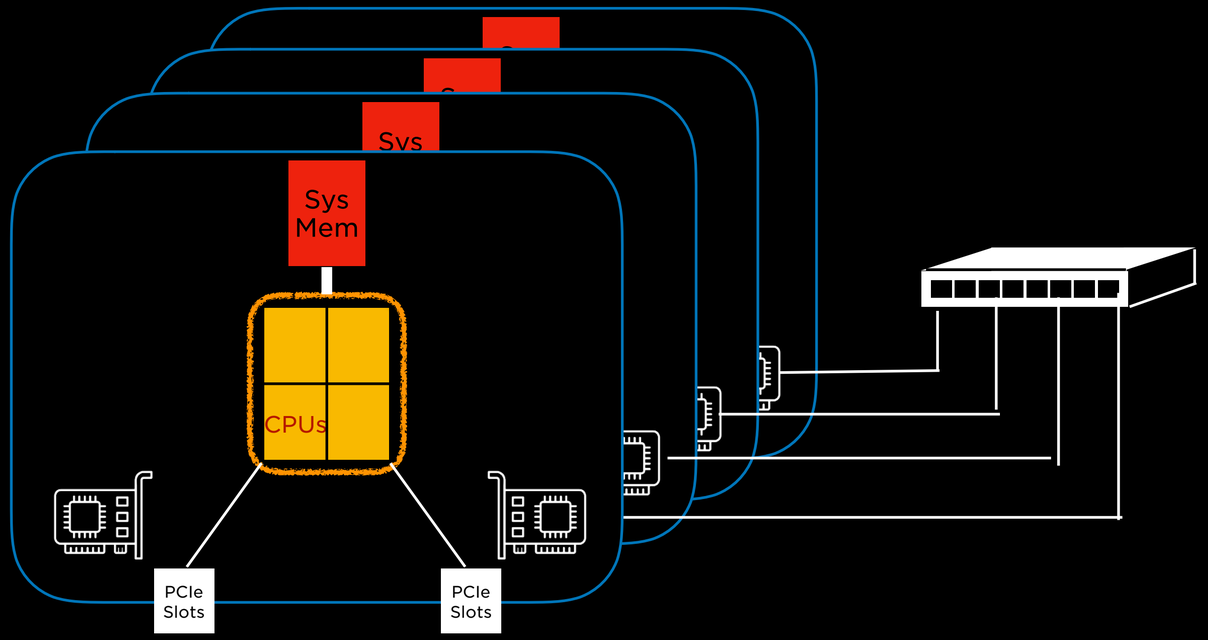
Integration Platform Constraints

Cooling Needs

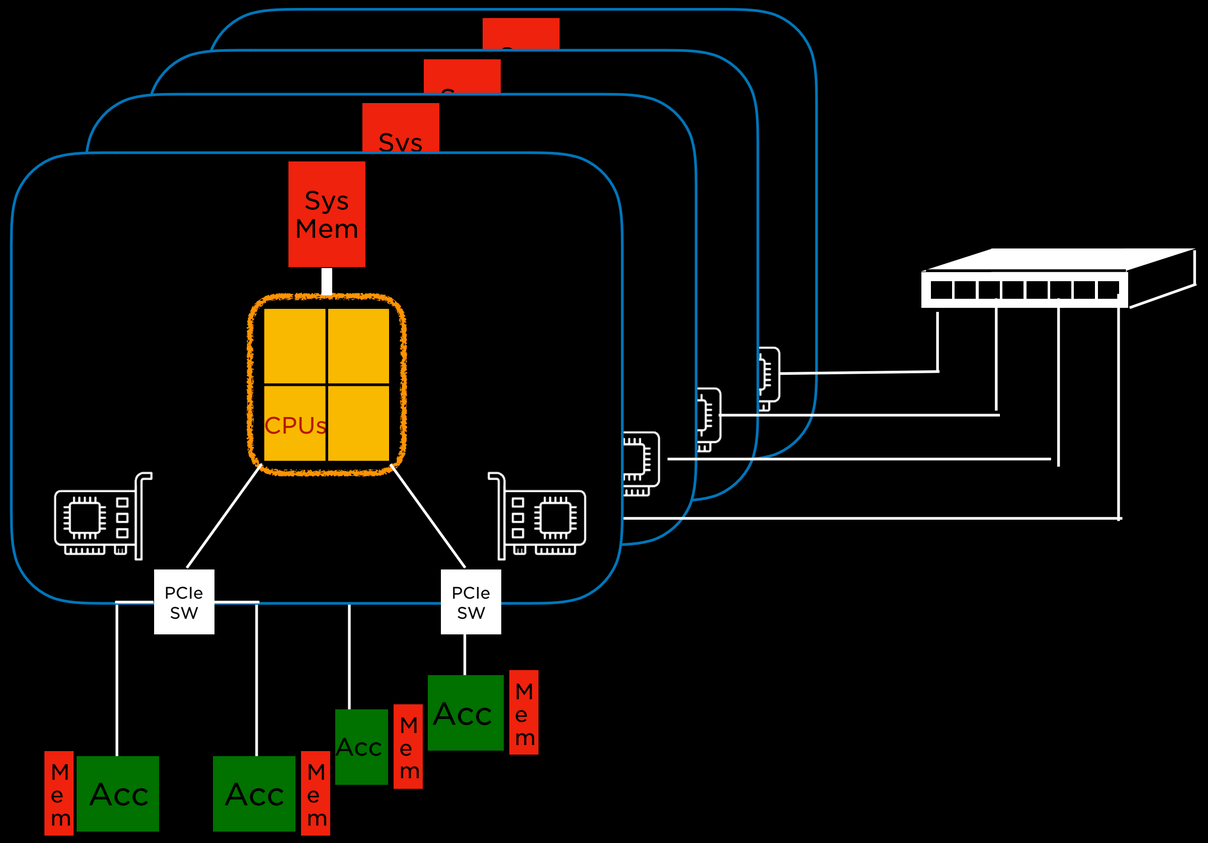
Power Delivery

Need Integrated Solutions With the Whole System in Mind !

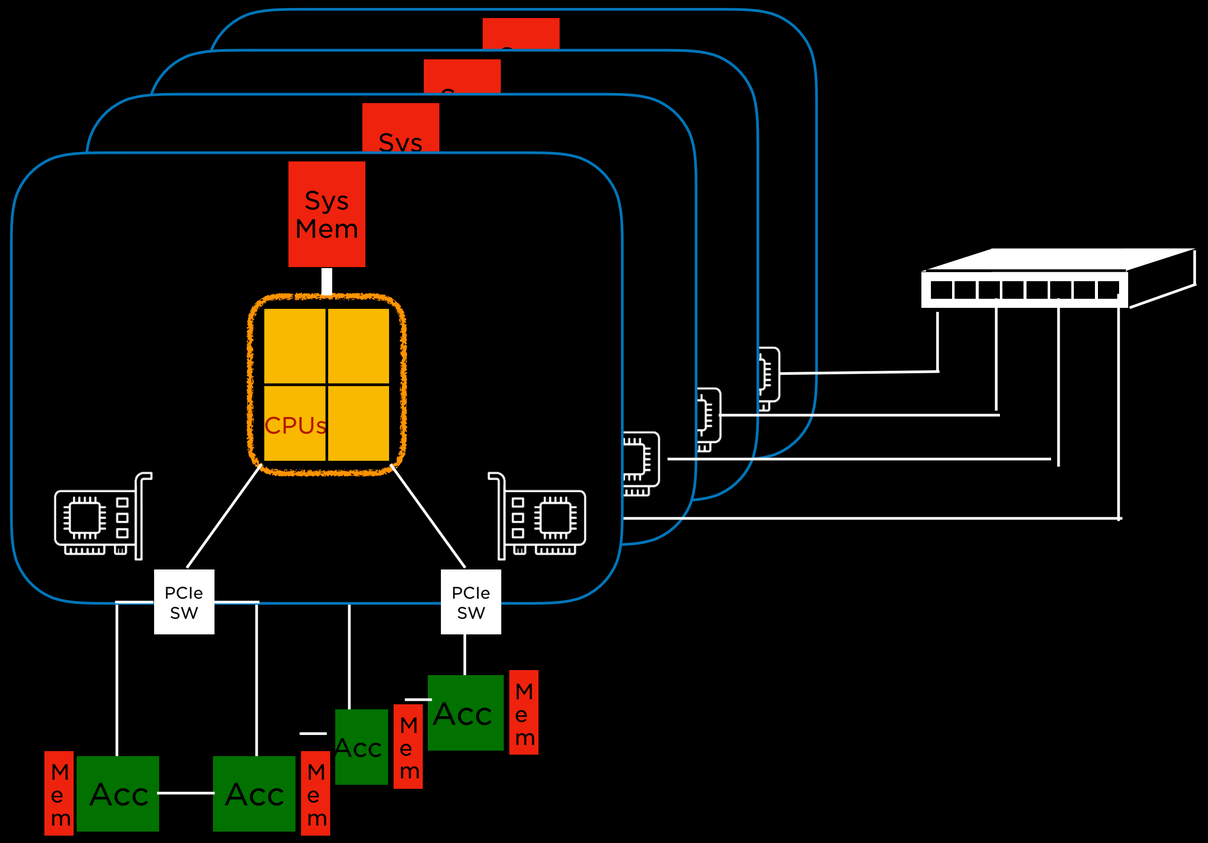
Datacenter Arch Evolution



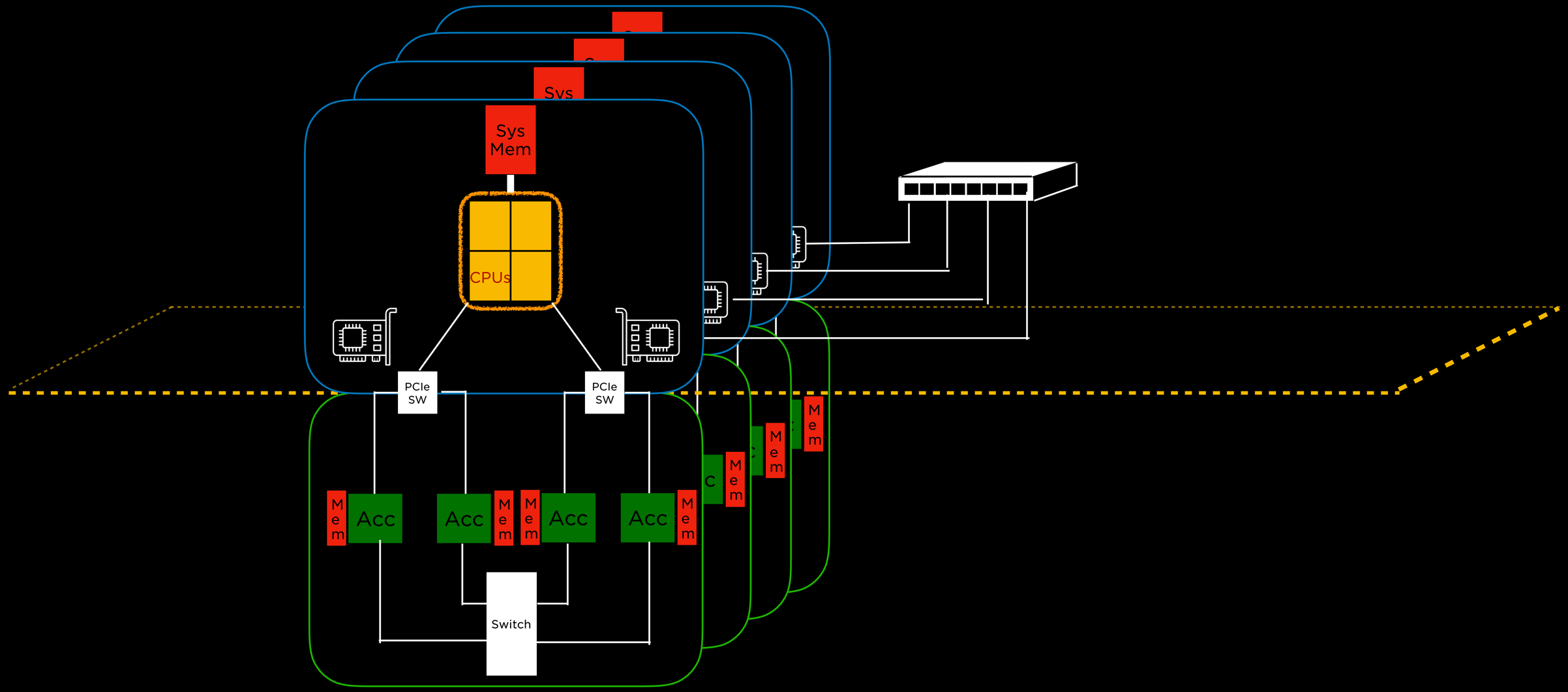
Datacenter Arch Evolution



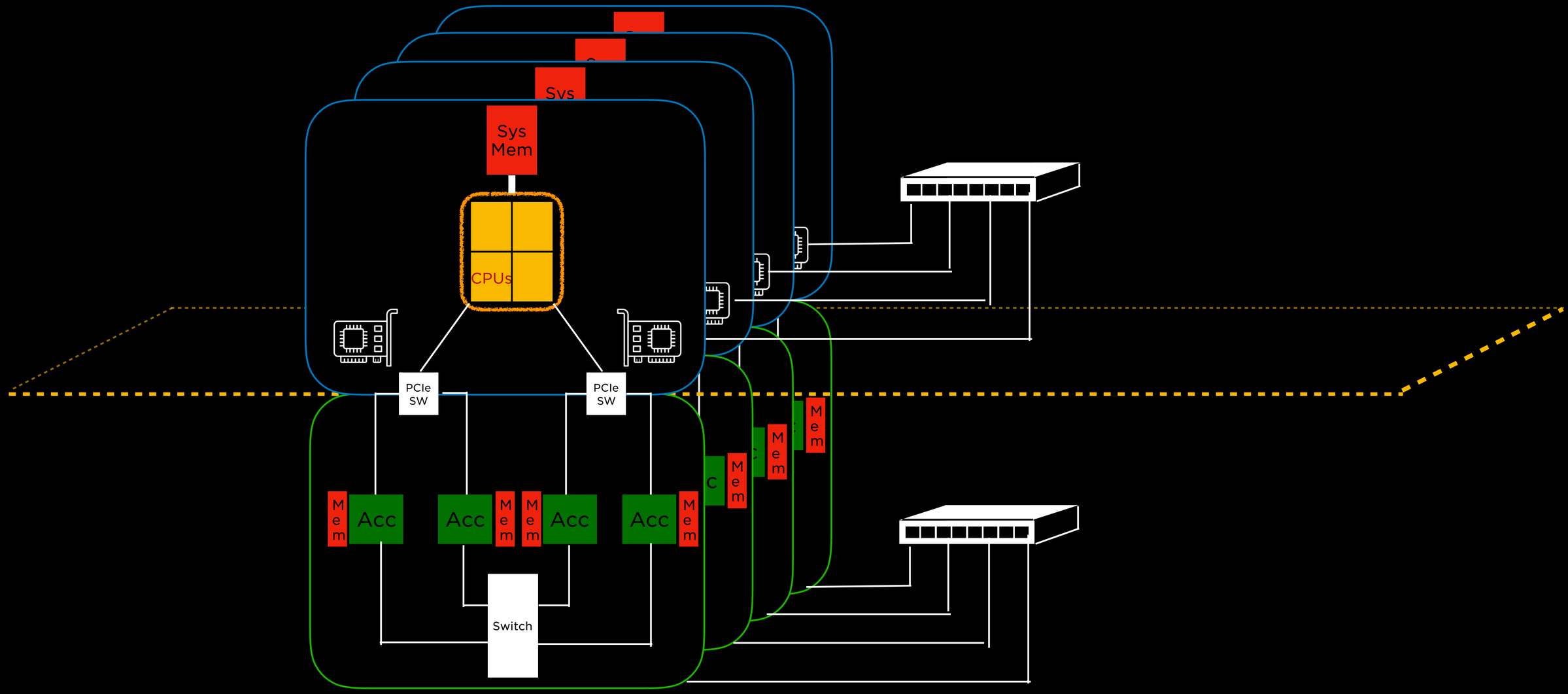
Datacenter Arch Evolution



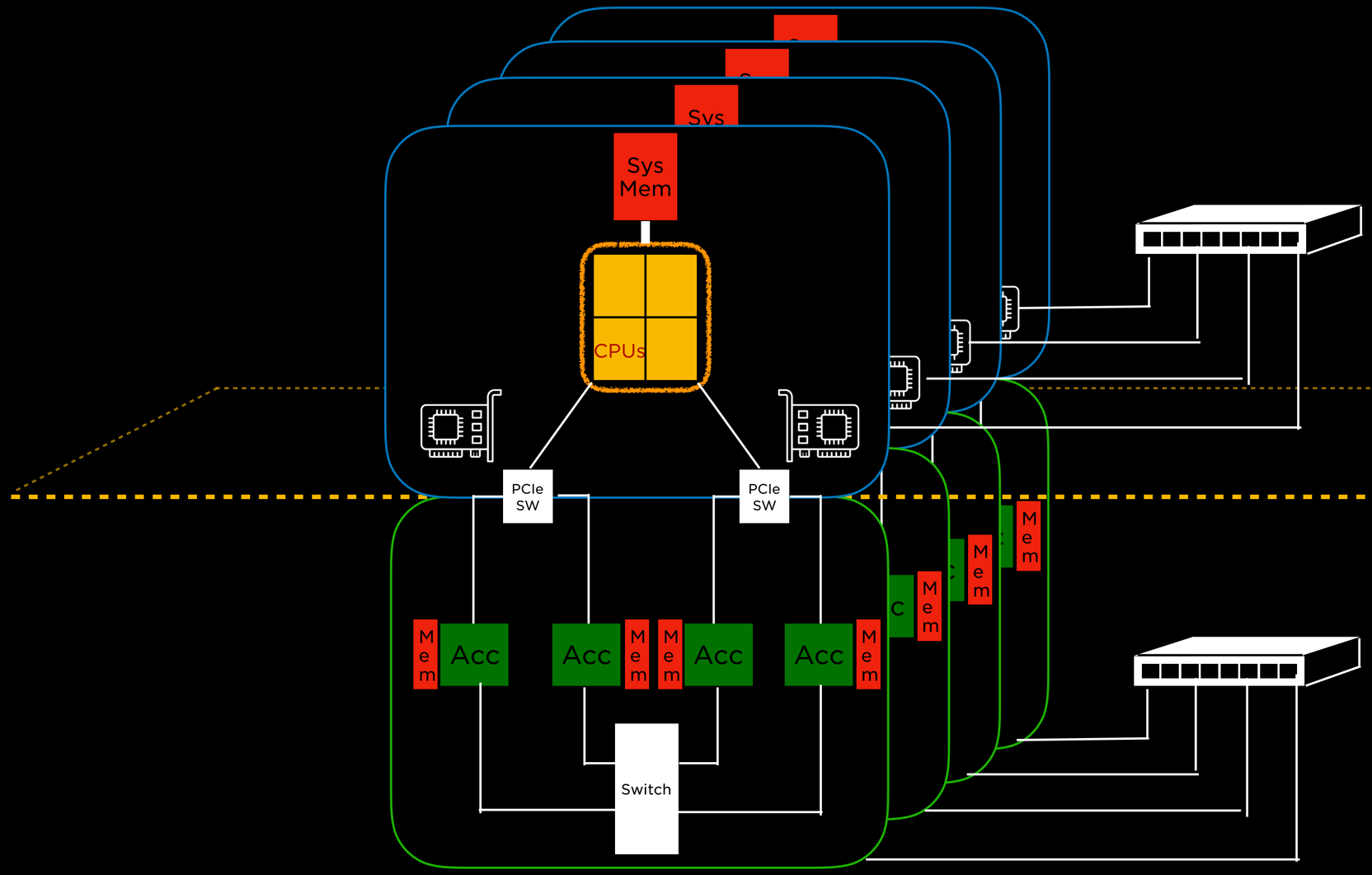
Datacenter Arch Evolution



Datacenter Arch Evolution



Datacenter Arch Evolution



Internet Optimized

Latencies ↑

Packet Sizes ↑

Bandwidth ↓

Legacy Support Baggage

HPC/CPU Centric

Rack-Rack Optimized

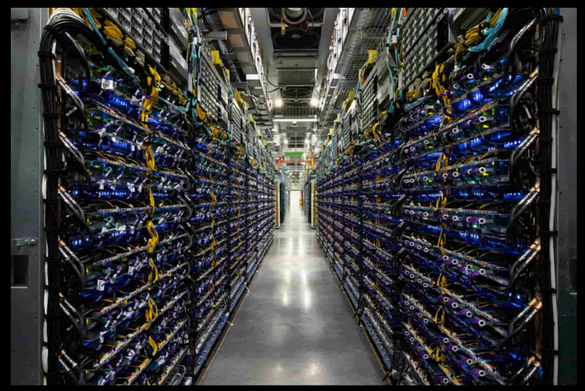
Latencies ↓

Bandwidth ↑

Packet Sizes ↓

AI Centric

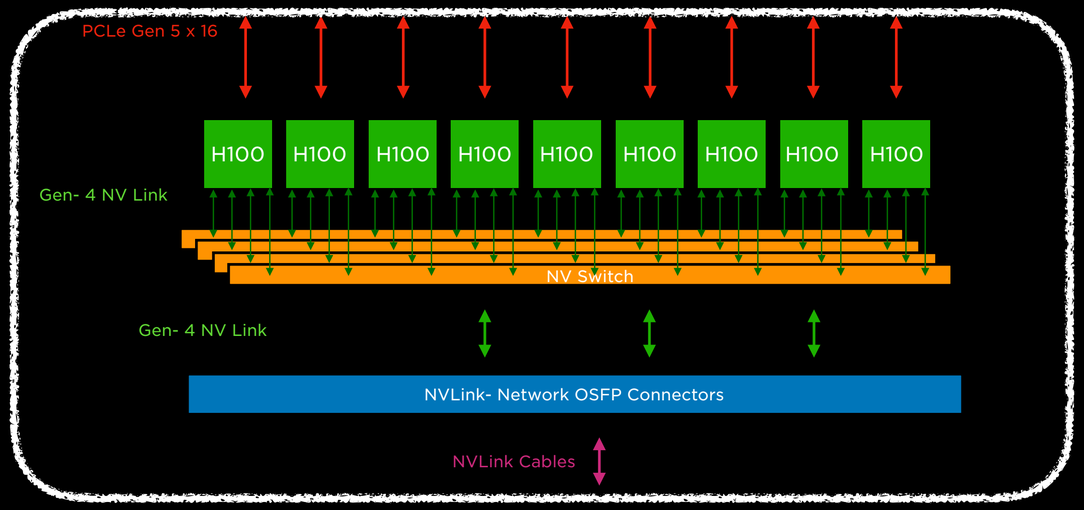
Beyond Compute - Communication Focus



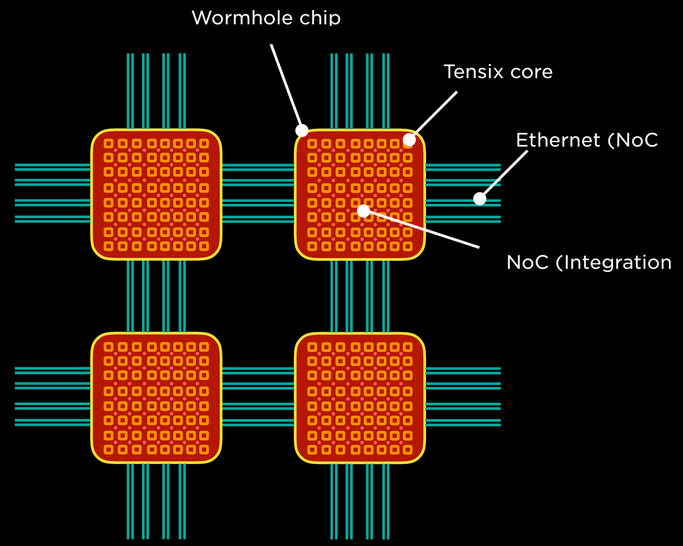
Source: cloud.google.com

Next Generation ML Infrastructure for large Model Training

TPU v4 chips are networked together into a Cloud TPU v4 pod by ultra-fast interconnect that provides 10x the bandwidth per chip at scale compared to typical GPU-based large scale training systems. Large models are very communication intensive: local computation often depends on results from remote computation that are communicated across the network. TPU v4's ultra-fast interconnect has an outsized impact on computational efficiency of large models by eliminating latency and congestion in the network.

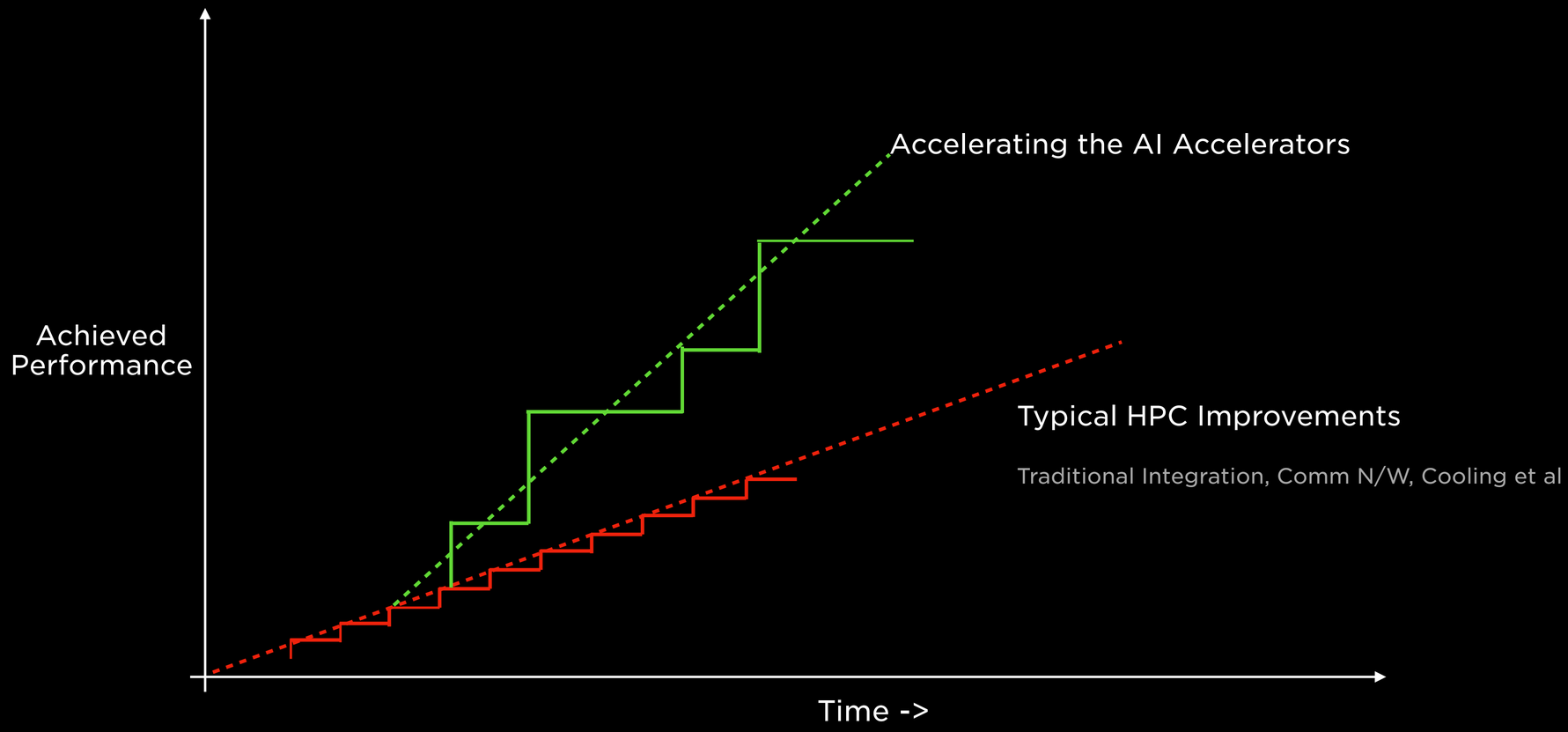


Source:https://developer.nvidia.com/



Source:Linley spring conference 2021

Rate of Change



Numerous Opportunities To Do Better/Quicker

Wider aperture beyond chips and into Systems

Reduce the Drag coefficient from the traditional hierarchies

Clean abstraction exists from frameworks to underlying HW

Flexible vs Fixed ratios of Compute-Memory-I/Os by disaggregation

Concentrate on the full solution stack

Can We Do Better ?

What would you do if designing from first principles for AI?

Dojo D1 Chip

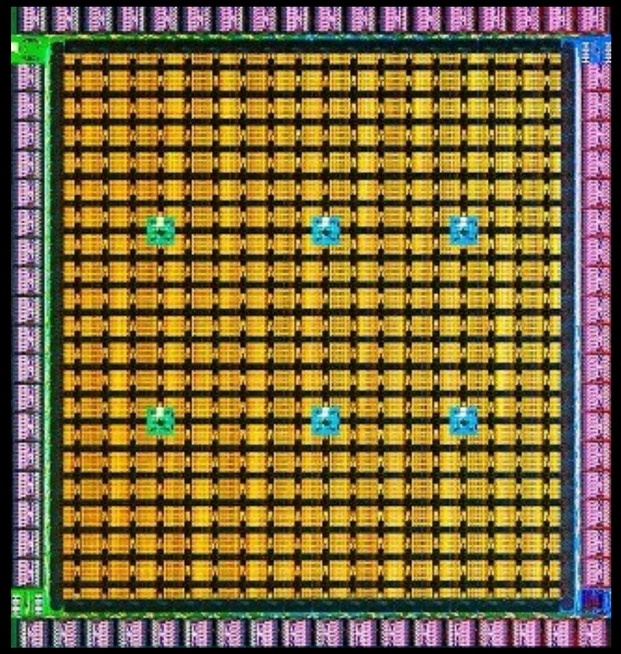
362 TFLOPs BF16/CFP8

22.6 TFLOPs FP32

10TBps/dir. On-Chip Bandwidth

4TBps/edge. Off-Chip Bandwidth

400W TDP

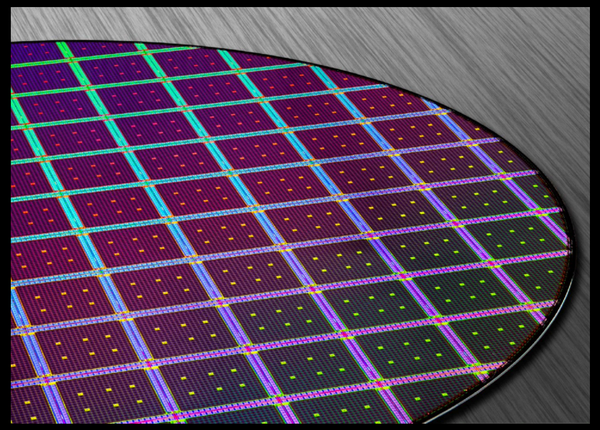


645mm²
7nm Technology

50 Billion
Transistors

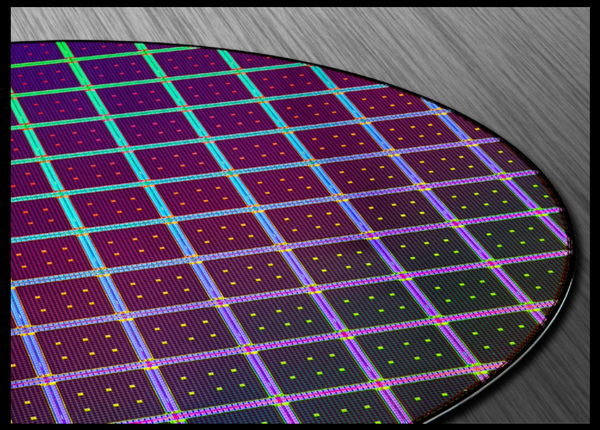
11+ Miles
Of Wires

Dojo Unique Innovation

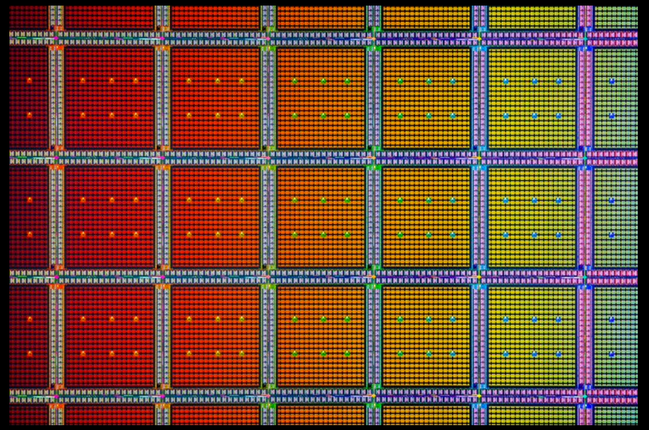


Silicon Wafer

Dojo Unique Innovation



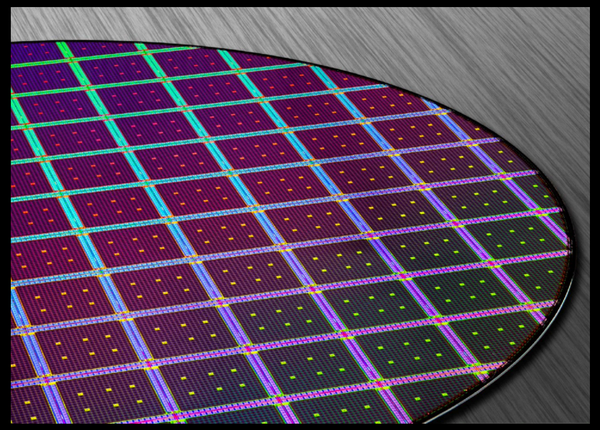
Test & Sort



Silicon Wafer

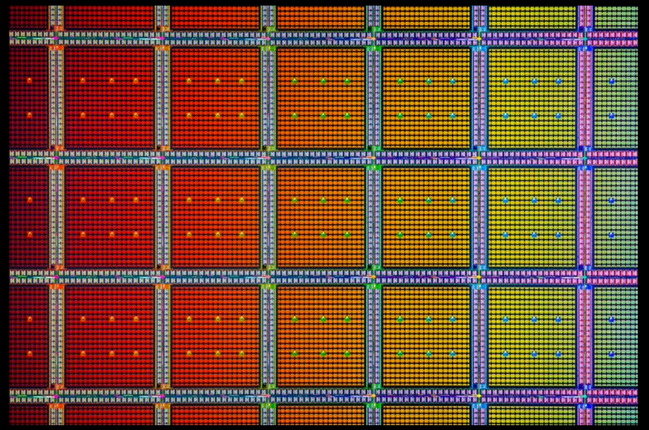
Known Good Dies

Dojo Unique Innovation



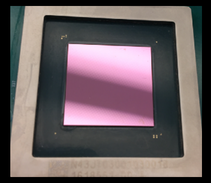
Silicon Wafer

Test & Sort

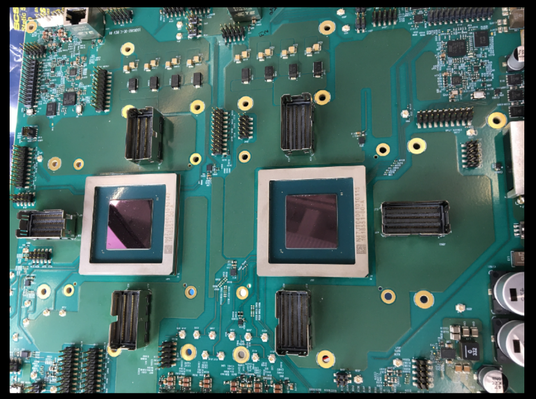


Known Good Dies

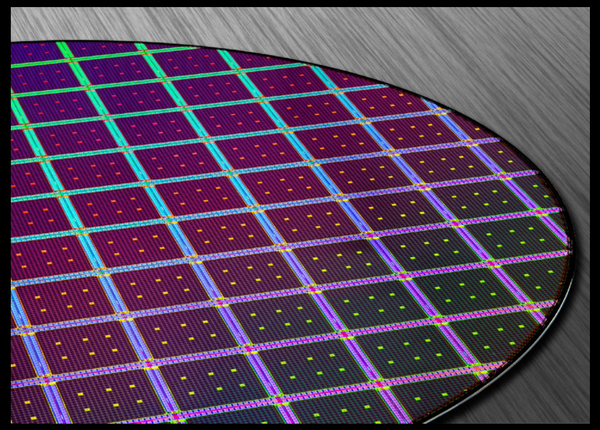
Pkg



PCB

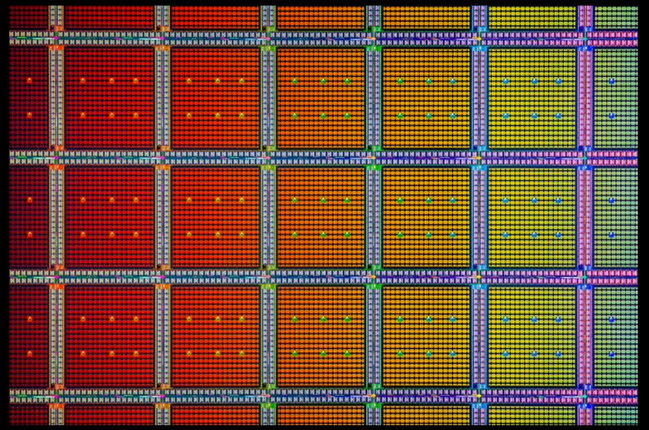


Dojo Unique Innovation: ~~Hierarchies~~



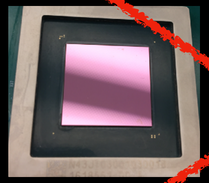
Silicon Wafer

Test & Sort

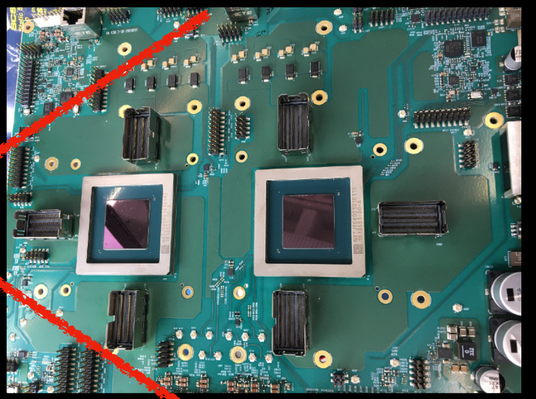


Known Good Dies

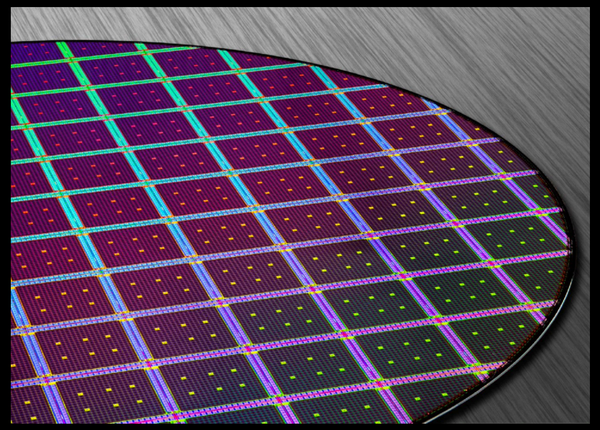
Pkg



PCB

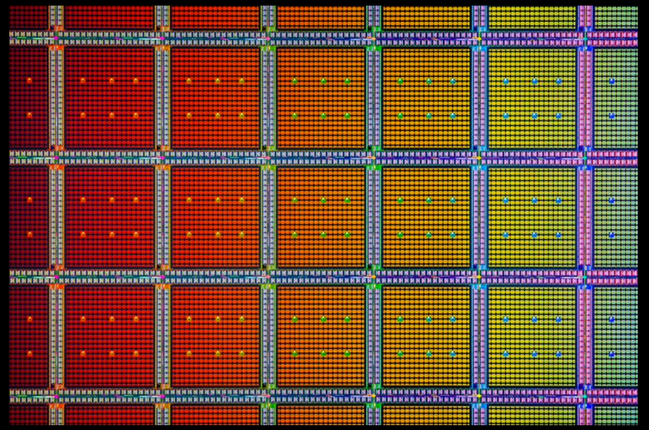


Dojo Unique Innovation: Flattened Hierarchies



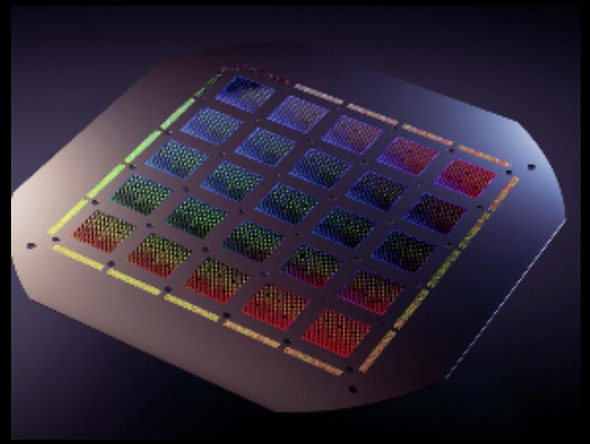
Silicon Wafer

Test & Sort



Known Good Dies

New Integration



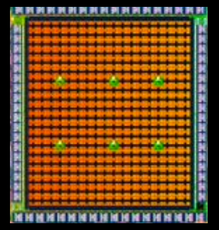
Reconstructed
Fanout Wafer

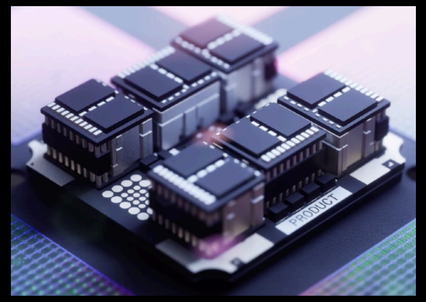
9 PFLOPs (BFP16/CFP8)

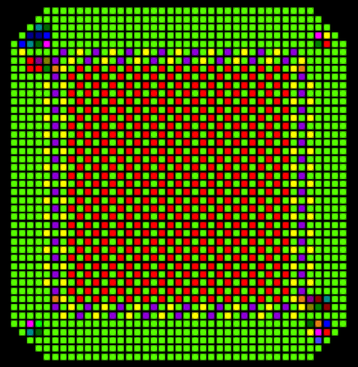
40 TB/s Bisec BW (X+Y)

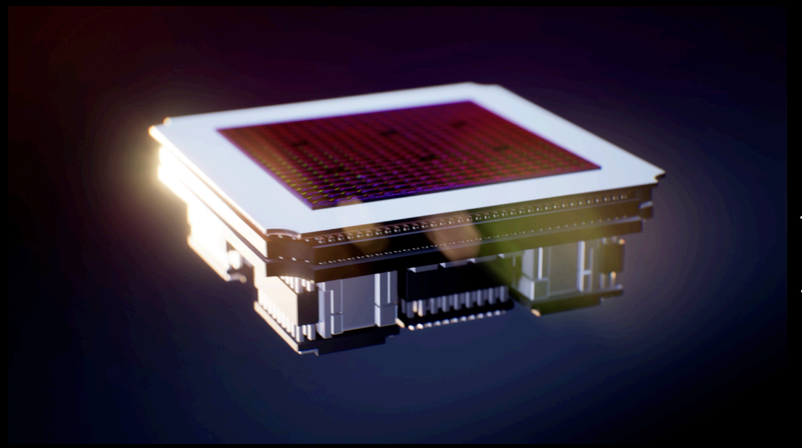
36TB/s I/O BW

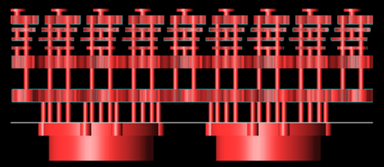
Vertical Power Delivery

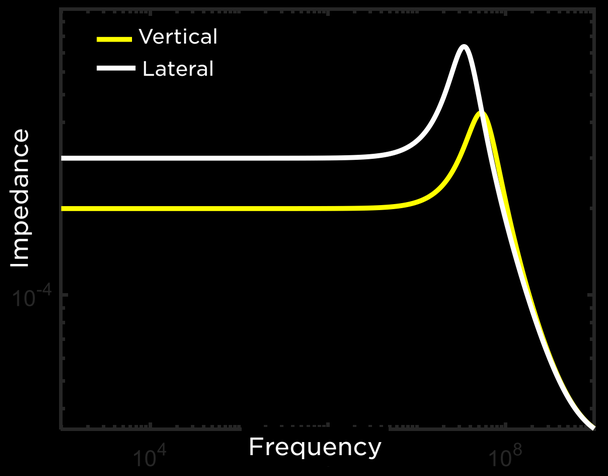




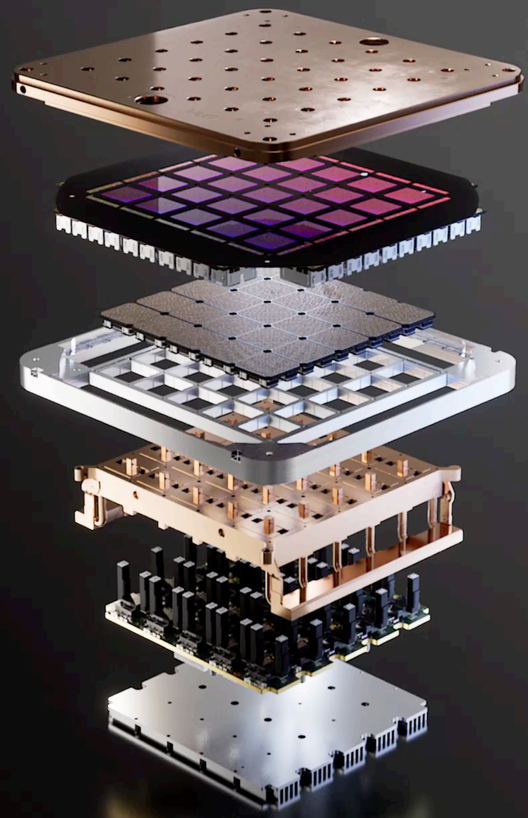








Training Tile



TESLA

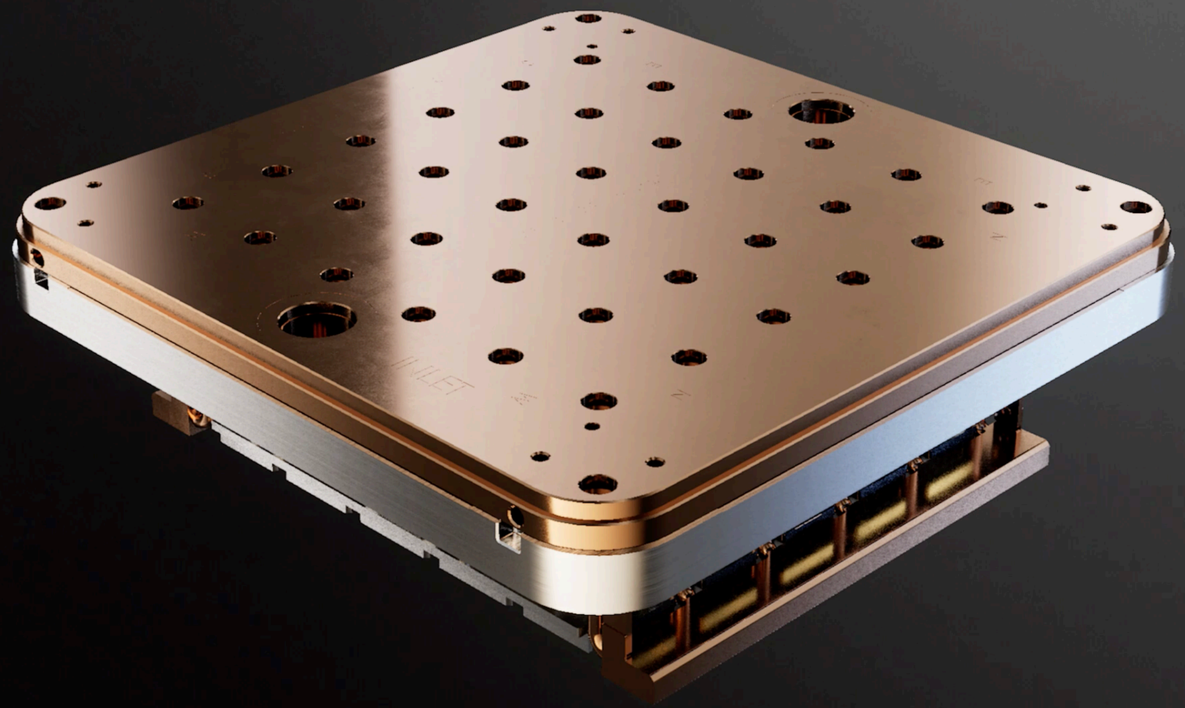
Training Tile

25x D1 Dies

9 PFLOPs

36TB/s I/O BW

< 0.5 cu Ft



High Compute Performance

Extremely High-Bandwidth

Low Latencies

Lower Energy Communication

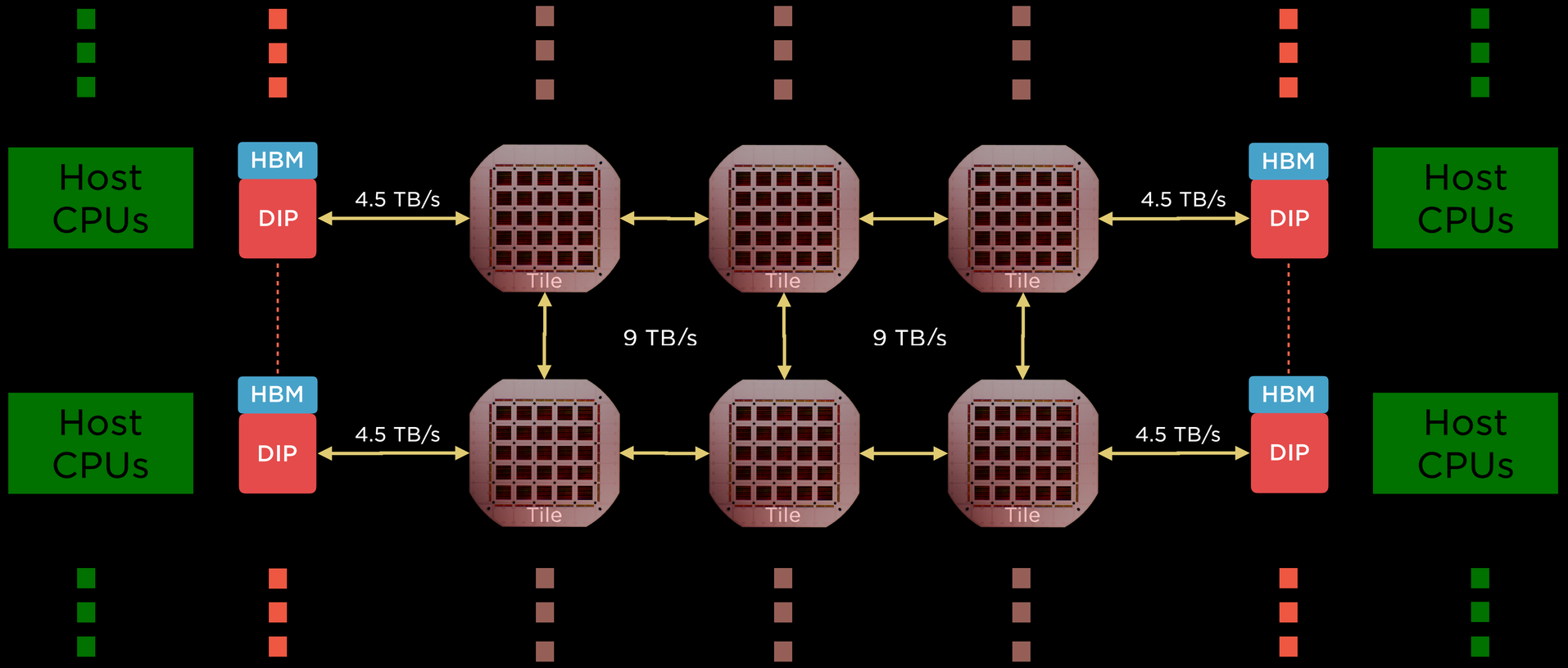
No external Switches

Unparalleled System Integration

Compute-Memory-Power Delivery-Cooling- Comm. ports

Tile (25 Chips) vs PCIe Card (1 or 2 Chips)

Flexible Dojo Training Matrix



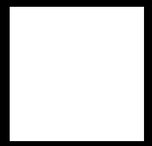
Fully Flexible Configurations

Feasible With Tile That Can Click Together in any Arrangement

Adaptable Ratios of Compute - Memory- I/O- Storage

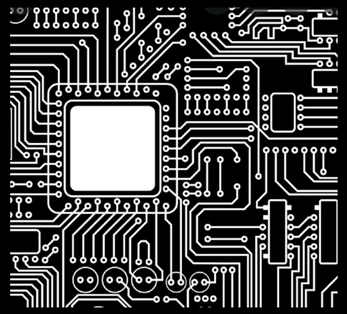
25 - 150 - 3000	D1 Dies
9 - 54 - 1080	PFLOPs BF16/CFP8
11 - 66 - 1320 GB	High-Speed SRAM
.16 - .64 - 13 TB	High-Bandwidth DRAM
4 - 8 - 80TB/s	Interface Low latency High BW Feeds

Traditional Hierarchies

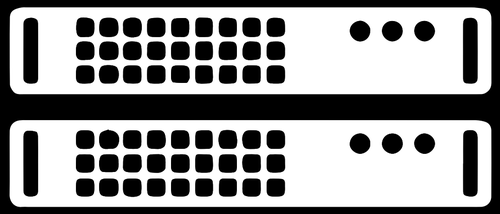


Chip

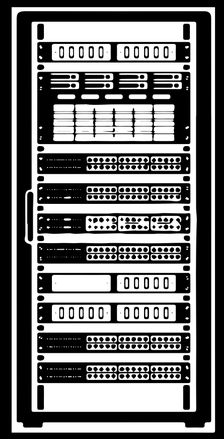
Package



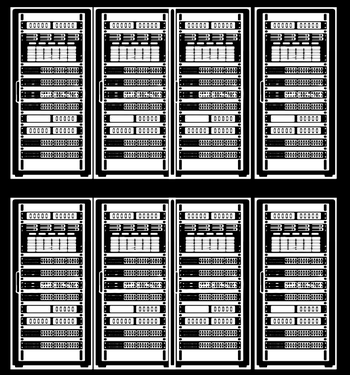
Boards



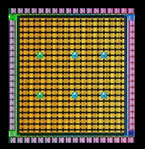
Boxes



Racks

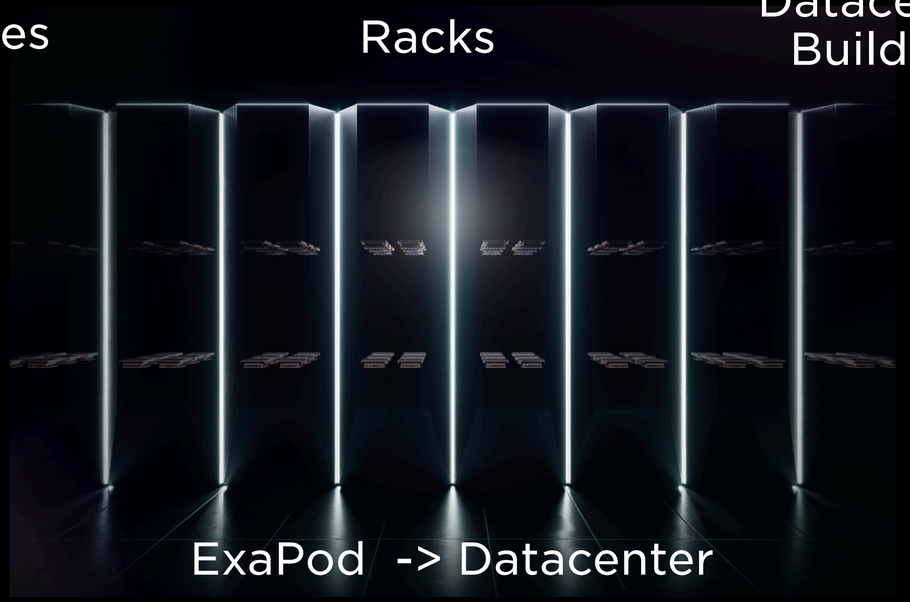


Datacenter/
Buildings



Chip

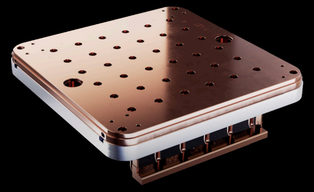
Tile



ExaPod -> Datacenter

Disaggregated Scalable System

Tile



Compute

Feeding the Beast(s)

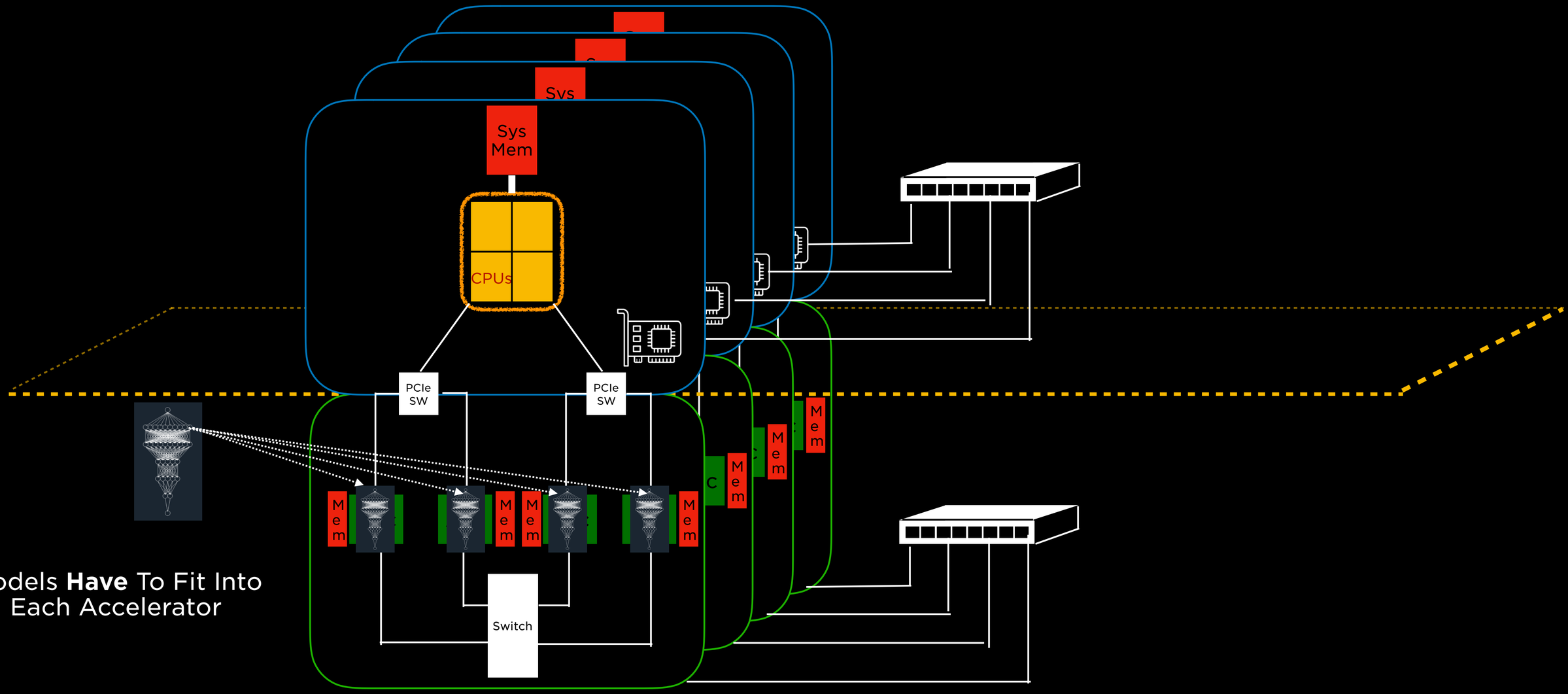
Interface Processor

Network Interface

Memory

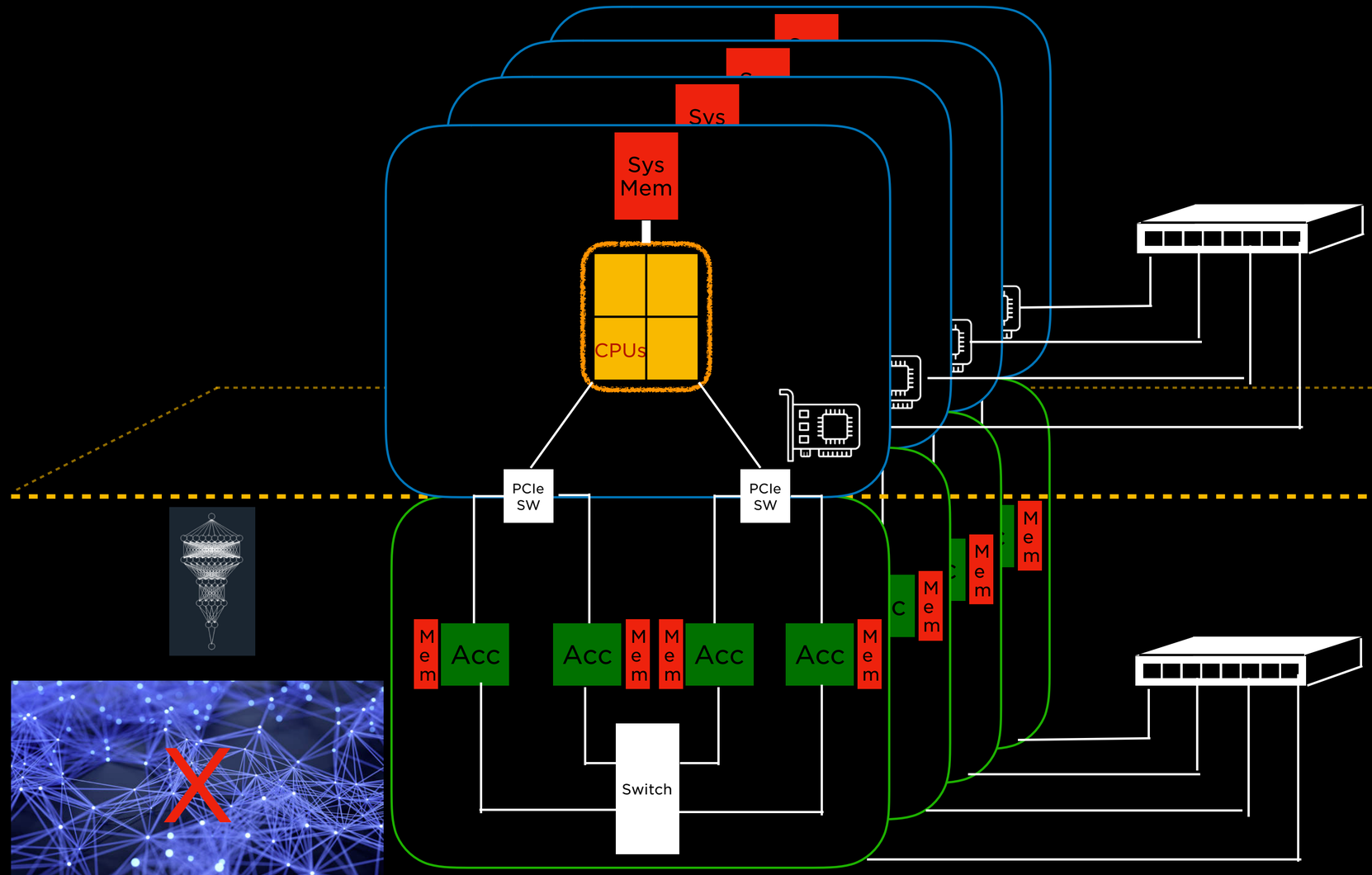
I/O

Traditional ML Model Fitting



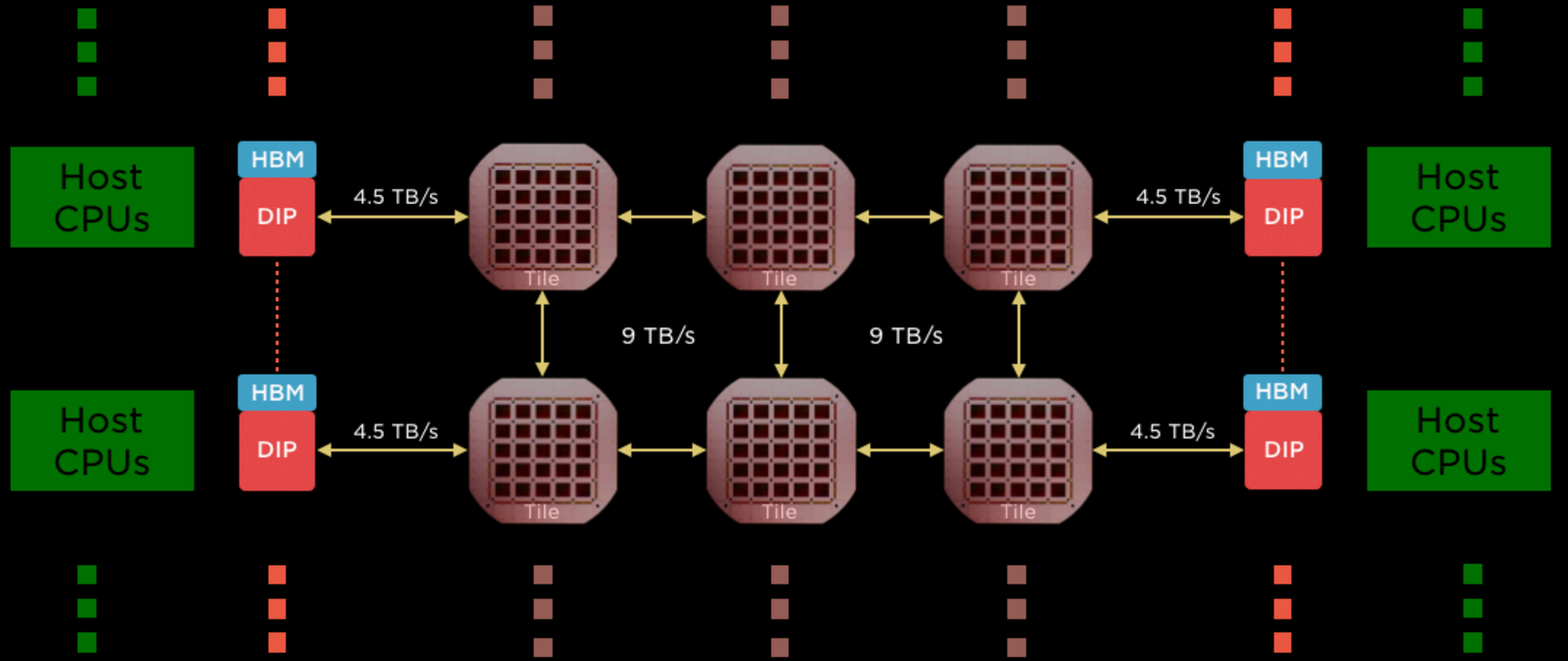
Models **Have** To Fit Into Each Accelerator

Traditional ML Gigantic Model Fitting Issues



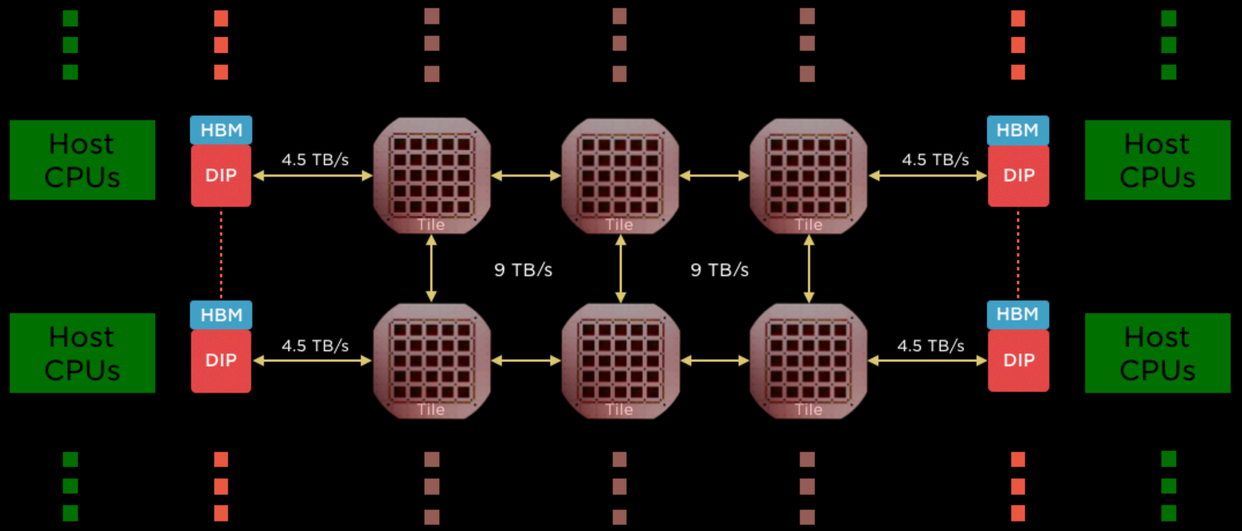
- Spilling to System Memory or Local Pool of Drams
- Can’t Assume Future Models Will Fit Into One Accelerator

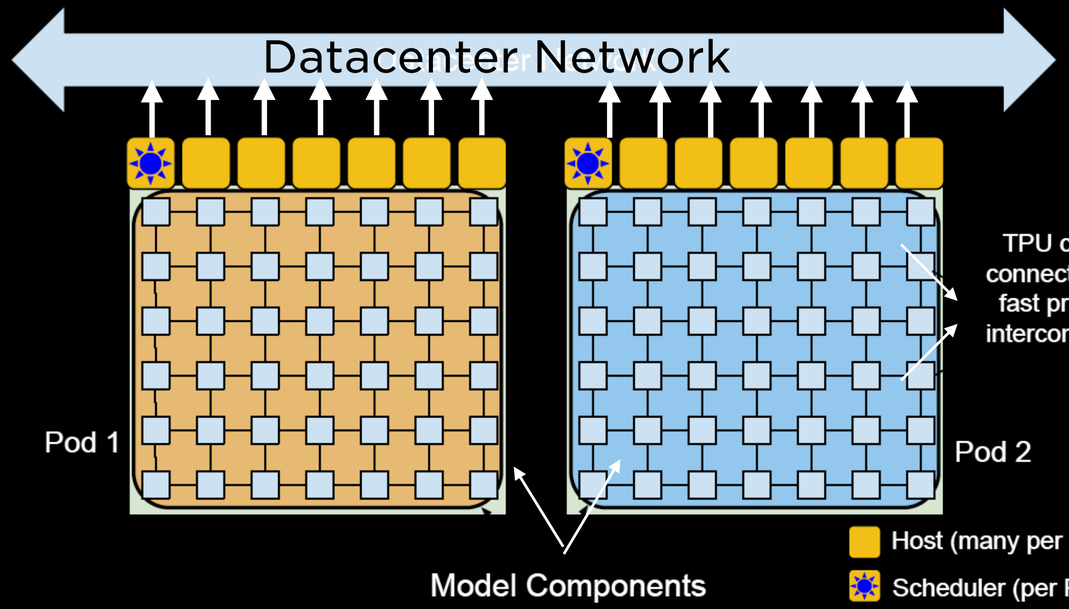
Gigantic Models



Designed from the get go to splitting Models across multiple Chips/Tiles

AI Focussed System Architectures

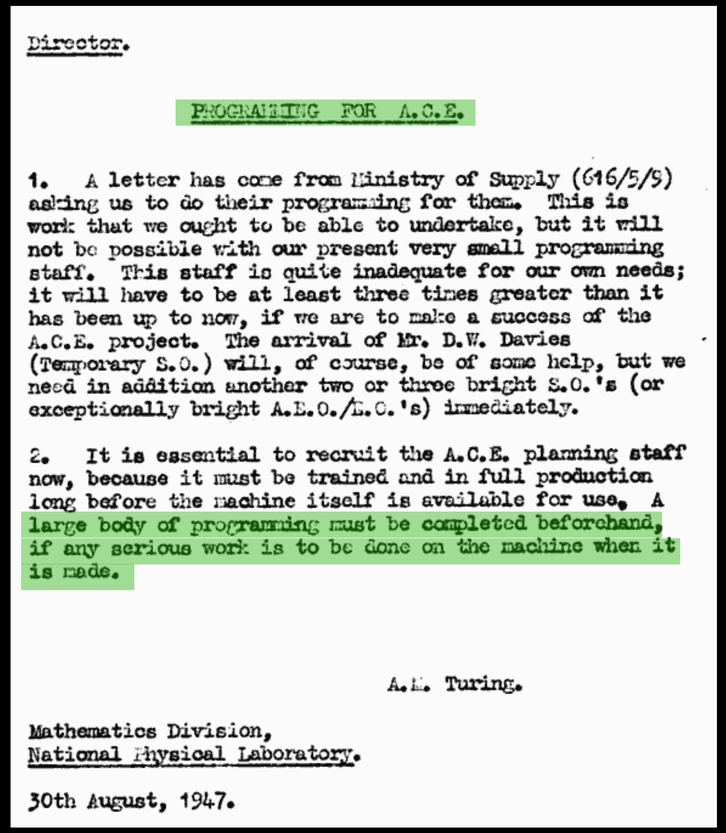




Source: Google Research: PaLM: Scaling Language Modeling with Pathways

Programmability & Flexibility

"A Large body of programming must be completed beforehand, If any serious work is to be done on the machine when it is made." - Alan Turing



Director.

PROGRAMMING FOR A.C.E.

1. A letter has come from Ministry of Supply (G16/5/S) asking us to do their programming for them. This is work that we ought to be able to undertake, but it will not be possible with our present very small programming staff. This staff is quite inadequate for our own needs; it will have to be at least three times greater than it has been up to now, if we are to make a success of the A.C.E. project. The arrival of Mr. D.W. Davies (Temporary S.O.) will, of course, be of some help, but we need in addition another two or three bright S.O.'s (or exceptionally bright A.E.O./E.O.'s) immediately.

2. It is essential to recruit the A.C.E. planning staff now, because it must be trained and in full production long before the machine itself is available for use. A large body of programming must be completed beforehand, if any serious work is to be done on the machine when it is made.

A.M. Turing.

Mathematics Division,
National Physical Laboratory.

30th August, 1947.

Source: http://www.alanturing.net/turing_archive/

Programmability & Flexibility

"A Large body of programming must be completed beforehand, If any serious work is to be done on the machine when it is made." - Alan Turing

Director.

PROGRAMMING FOR A.C.E.

1. A letter has come from Ministry of Supply (G16/5/S) asking us to do their programming for them. This is work that we ought to be able to undertake, but it will not be possible with our present very small programming staff. This staff is quite inadequate for our own needs; it will have to be at least three times greater than it has been up to now, if we are to make a success of the A.C.E. project. The arrival of Mr. D.W. Davies (Temporary S.O.) will, of course, be of some help, but we need in addition another two or three bright S.O.'s (or exceptionally bright A.E.O./E.O.'s) immediately.

2. It is essential to recruit the A.C.E. planning staff now, because it must be trained and in full production long before the machine itself is available for use. A large body of programming must be completed beforehand, if any serious work is to be done on the machine when it is made.

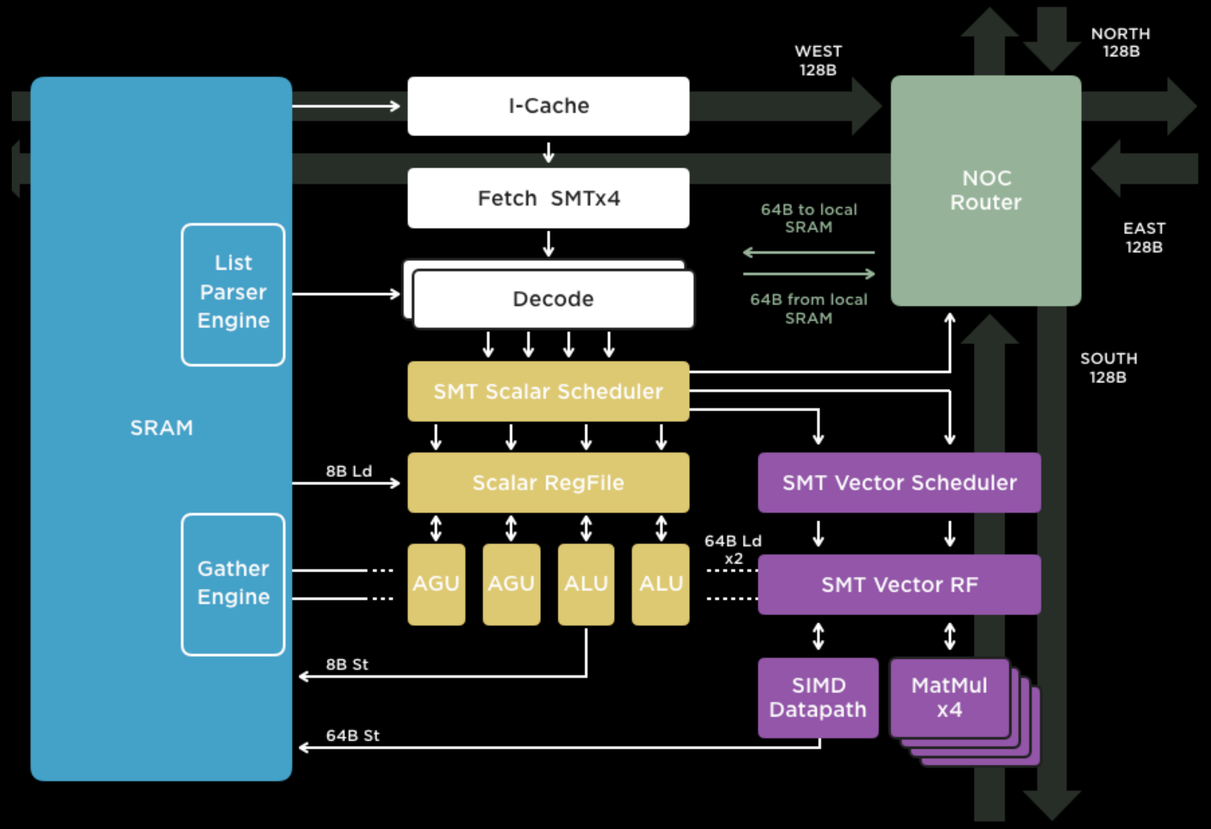
A.M. Turing.

Mathematics Division,
National Physical Laboratory.

30th August, 1947.

Source: http://www.alanturing.net/turing_archive/

Ease of programability is essential :



Fully Flexible, Compiler Friendly Yet High Perf Architectures

Software Stack

Neural Net Models

PyTorch

Dojo Extension

PyTorch-Extension

Dojo Compiler Engine

LLVM

JIT NN Compiler

LLVM Backend

Dojo Drivers

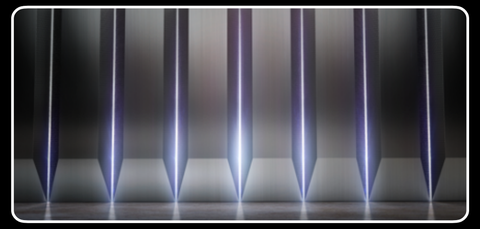
Multi-Host,
Multi Partition Management

PCIe

Dojo Interface Processors

Ingest & Shared Mem

Serdes



ExaPOD

HW Help for SW Stacks

Compiler Friendly ISAs

HW Sync/Barriers

Flexible StateMachines for ML Layers

Fire and Forget Communication Protocols

Fault Tolerance

Software Stack Opportunities

Take Advantage of Clean Abstraction Layers for ML

Need Changes for Massively Parallel Architectures

More Compilers, Less Kernels

Need Renewed Focus on Distributed Compiler Technology

Reduced OS Roles

Beyond Compute: Many Aspects - One Goal

DataTypes : CFP Formats for Efficiency

Compute : CPU+GPU+NPU+NNA

Scaleout : Seamless Scale Out Fabric

Communication : TerraBytes per Sec

Memory : GBs/TBs

Disaggregation : Ratios Move With Workloads

Compiler Technologies : Truly Distributed Compilers

Networking Topologies : TTP/TTPoEs

Framework Enhancements : Dojo-PyTorch

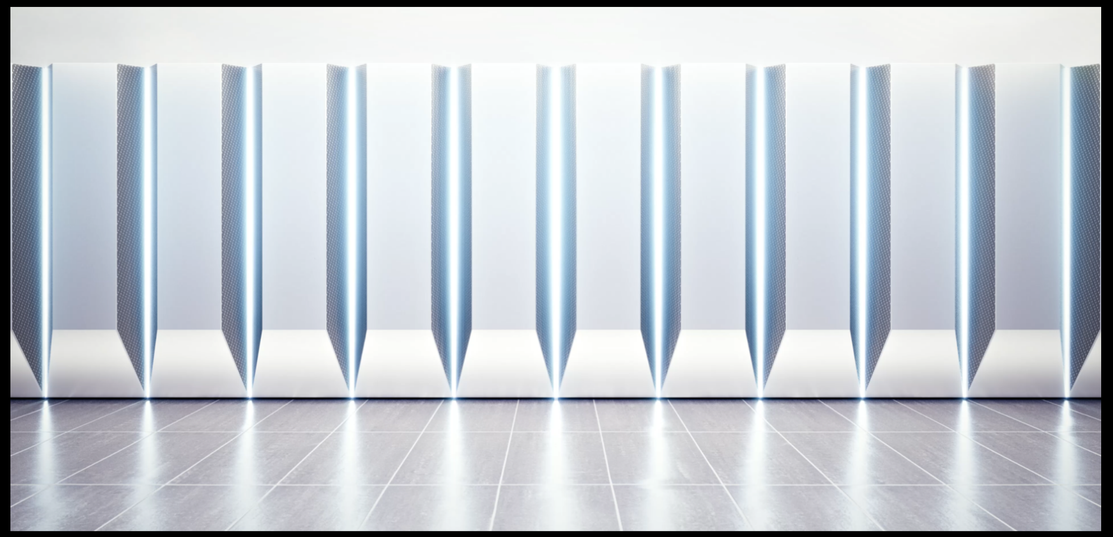
.....

...

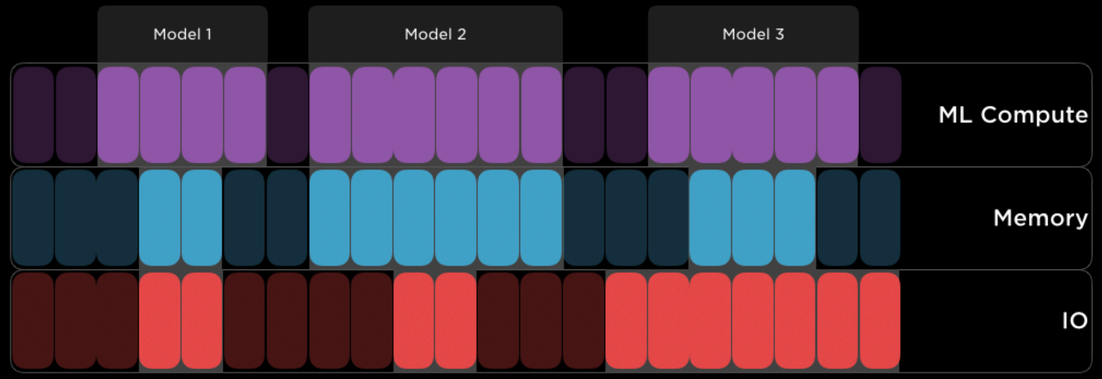
.

Beyond Compute: Scaleout From ML to AI !

Tightly Integrated Yet Disaggregated System



ExaPod



Job Specific Sizing of Resources

Innovation Opportunities Beyond Compute for AI Systems

Architectures :	Scale Out Focused Parallel Architectures
Integration :	Reduce Traditional Hierarchy Imposed Tax on Performance and Power
Disaggregation :	Alterable Ratios of Compute/Memory/Comm./Storage
Abstractions :	Taking Advantage of Clean Abstraction Layers of Frameworks
Algorithms :	Flexibility of Compute To Adapt to New Algorithms and Workloads
Compilers :	Explore/Revive Distributed Compiler Technologies

Innovation Opportunities Beyond Compute for AI Systems

Architectures : Scale Out Focused Parallel HW Architectures

Integration : Reduce Traditional Hierarchy Imposed Tax on Performance and Power

Disaggregation : Alterable Ratios of Compute/Memory/Comm./Storage

Abstractions : Taking Advantage of Clean Abstraction Layers of Frameworks

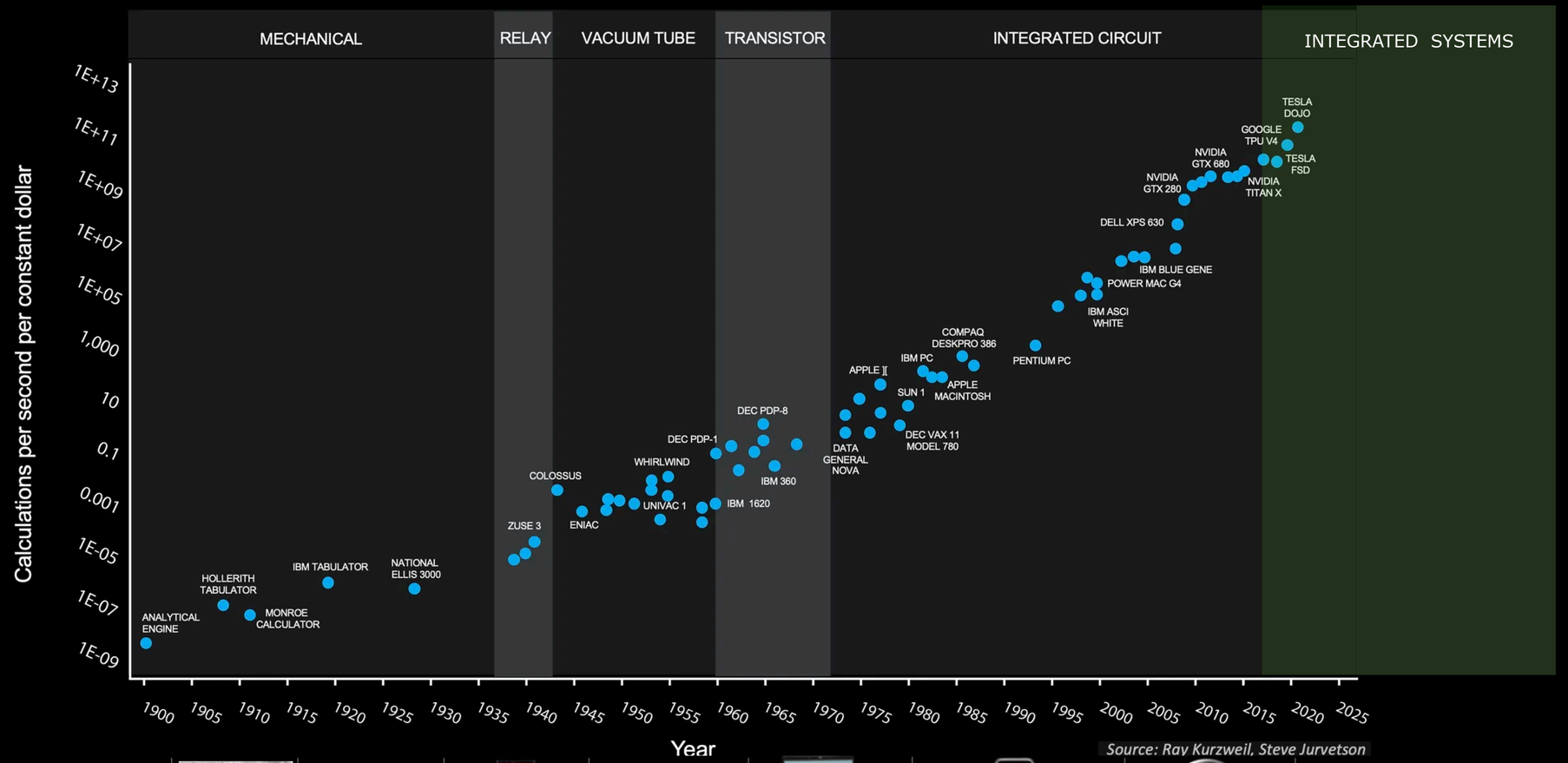
Algorithms : Flexibility of Compute To Adapt to New Algorithms and Workloads

Compilers : Explore/Revive Distributed Compiler Technologies

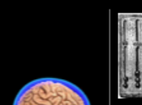
Design Approach of Exploring the Full Solution Space Across System and Software

Next Phase in Computing Evolution

122 Years of Moore's Law



Source: Ray Kurzweil, Steve Jurvetson


Human Computer


Abacus


Diff Engine


ENIAC


Calculator


Personal Computer


Cray-1


Laptop Computer


Consoles

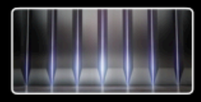

Smart Phone Computer


Datacenters

IoT Computer


FSD Computer


Training Server


Training Datacenters

Not on Same Scale
As the Above Chart

This is only a foretaste of what is to come, and only the shadow of what is going to be. We have to have some experience with the machine before we really know its capabilities . . . I do not see why it should not enter any one of the fields normally covered by the human intellect, and eventually compete on equal terms.

Alan Turing

(Quoted in *The Times*, 11 June 1949: 'The Mechanical Brain')

Thank You!

Learn more at

Tesla.com/AI