The Future of Computing Beyond Exascale

Looks a lot more like the past 30 years of wide-area distributed services

John Shalf

Department Head for Computer Science

Lawrence Berkeley National Laboratory

December 9, 2022

Celebration of (and with) Cees de Laat

University of Amsterdam (UvA)

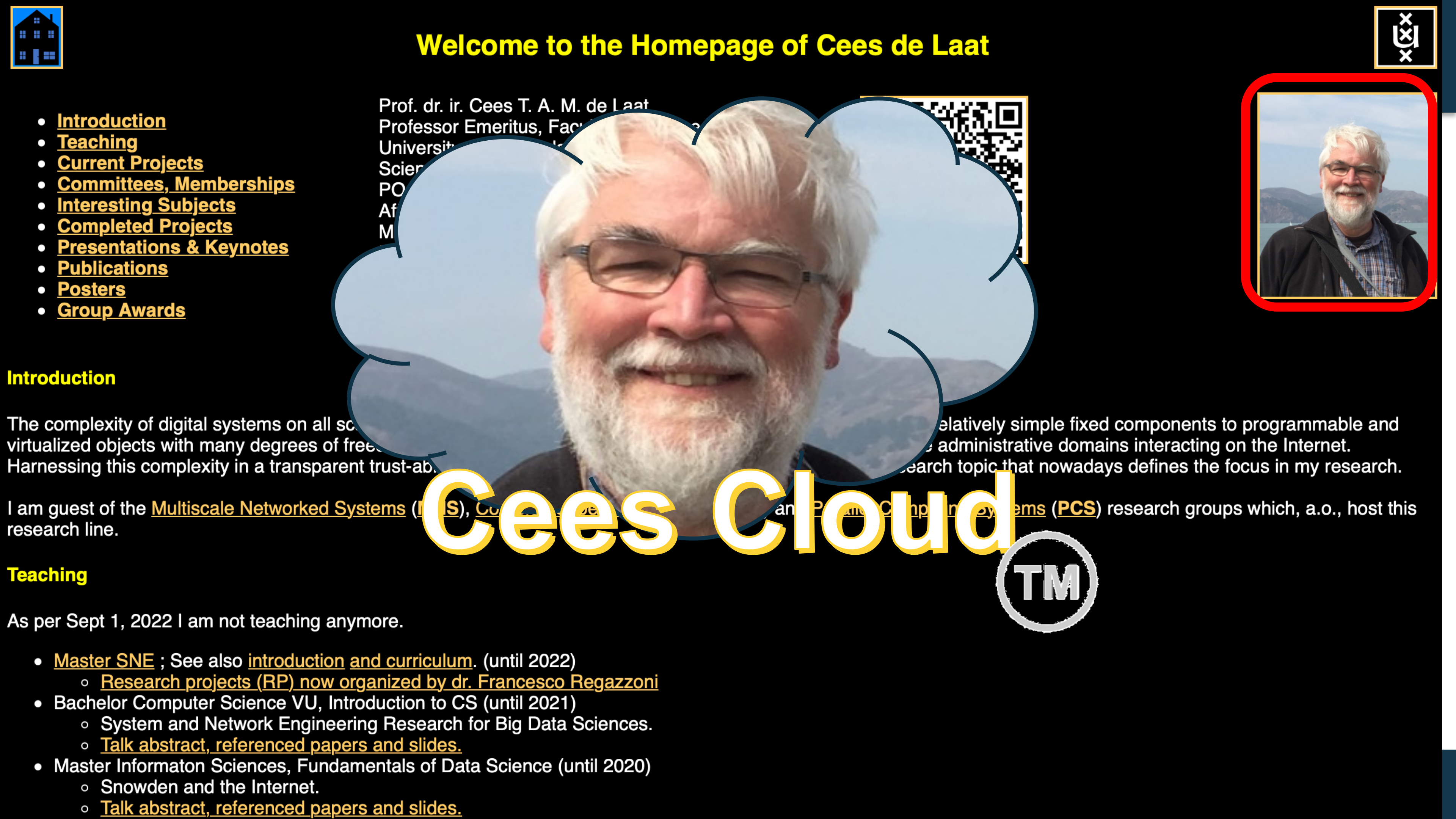
Last time I was in Amsterdam (October 2019)

Jun Xiao's
PhD Committee

me

Christopher Columbus

COFFEESHOP
COFFEESHOP
The Greenhouse Effe
Open



Welcome to the Homepage of Cees de Laat

- Introduction
- Teaching
- Current Projects
- Committees, Memberships
- Interesting Subjects
- Completed Projects
- Presentations & Keynotes
- Publications
- Posters
- Group Awards

Prof. dr. ir. Cees T. A. M. de Laat
Professor Emeritus, Fac
Universit
Scien
PO
Af
M

Introduction

The complexity of digital systems on all sc … elatively simple fixed components to programmable and virtualized objects with many degrees of free … administrative domains interacting on the Internet. Harnessing this complexity in a transparent trust-ab … earch topic that nowadays defines the focus in my research.

I am guest of the Multiscale Networked Systems (MNS), Co … an Parallel Computing Systems (PCS) research groups which, a.o., host this research line.

Teaching

As per Sept 1, 2022 I am not teaching anymore.

- Master SNE ; See also introduction and curriculum. (until 2022)
  - Research projects (RP) now organized by dr. Francesco Regazzoni
- Bachelor Computer Science VU, Introduction to CS (until 2021)
  - System and Network Engineering Research for Big Data Sciences.
  - Talk abstract, referenced papers and slides.
- Master Informaton Sciences, Fundamentals of Data Science (until 2020)
  - Snowden and the Internet.
  - Talk abstract, referenced papers and slides.

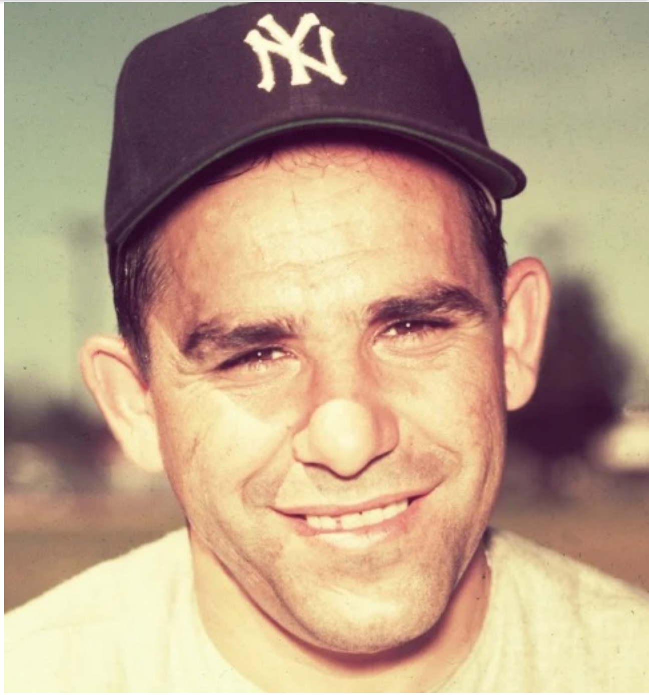
The Future of post-exascale HPC is looking more like the past 30 years of High Performance Distributed Services/Networking

Its like Deja-Vu all over again!

- *Yogi Berra*

You can observe a lot just by watching

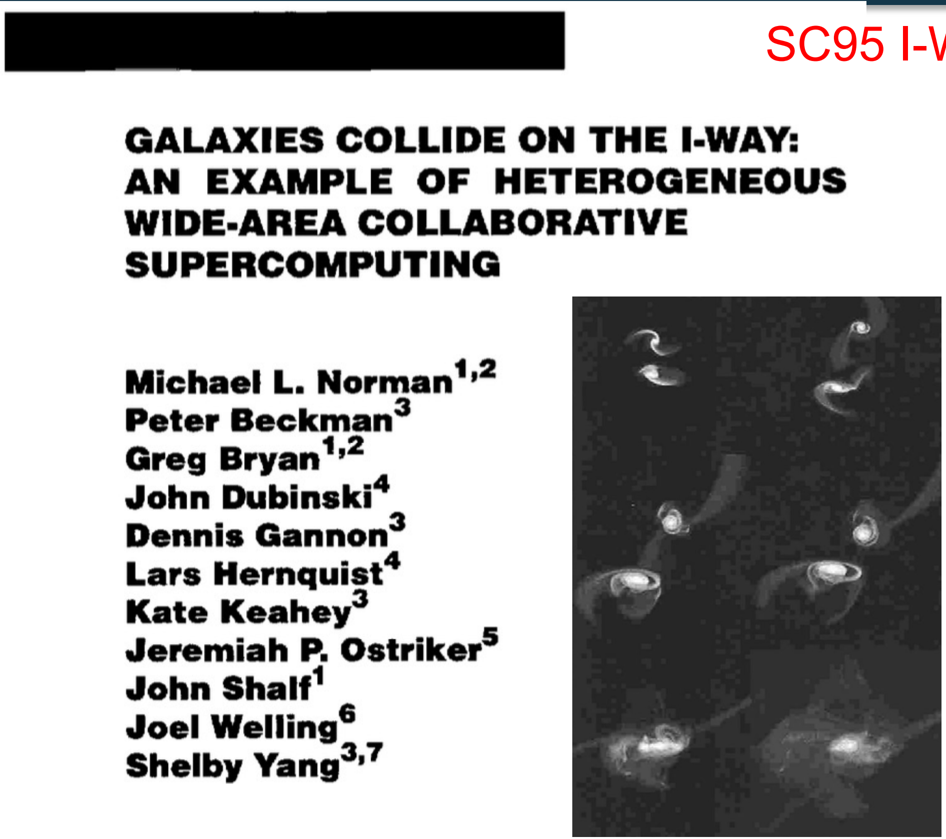
- *Same guy*



Yogi Berra

My Past Foray into High Performance Distributed Computing

[redacted]



GALAXIES COLLIDE ON THE I-WAY: AN EXAMPLE OF HETEROGENEOUS WIDE-AREA COLLABORATIVE SUPERCOMPUTING

Michael L. Norman[1,2]
Peter Beckman[3]
Greg Bryan[1,2]
John Dubinski[4]
Dennis Gannon[3]
Lars Hernquist[4]
Kate Keahey[3]
Jeremiah P. Ostriker[5]
John Shalf[1]
Joel Welling[6]
Shelby Yang[3,7]

SC95 I-WAY

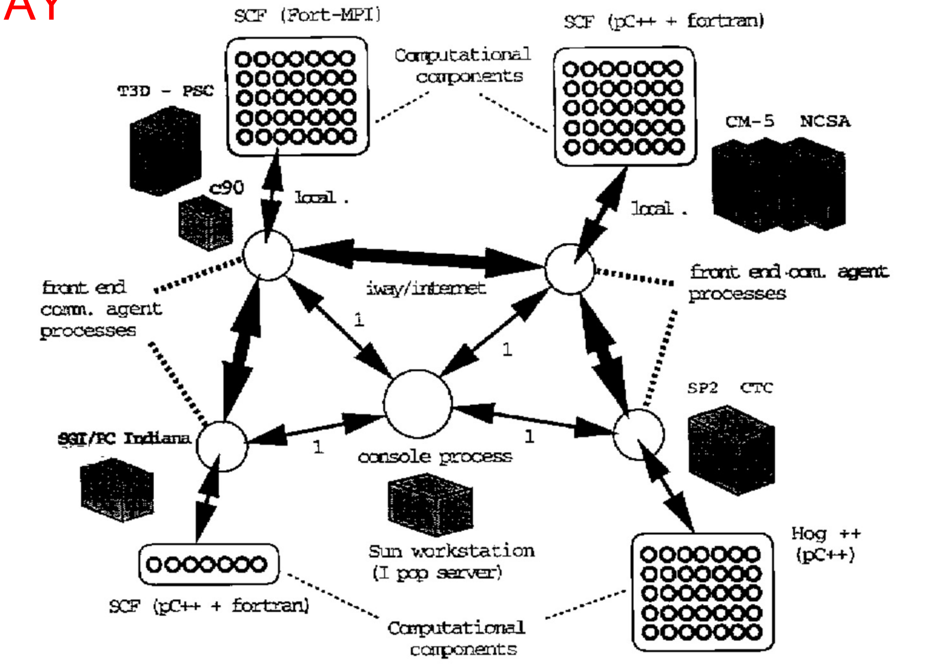


Fig. 2 Architecture of our distributed heterogeneous I-WAY application

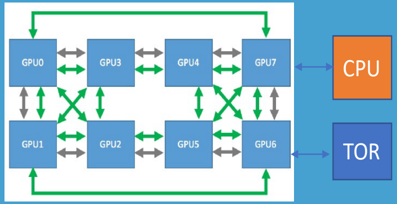
My Past Foray into High Performance Distributed Computing

- ***Then:*** **StarTap (1997) followed by StarLIGHT and NetherLight and GLIF (2001)** *(working for Tom Defanti)*
  - **Emerging Global Movement:** Eighth Joint European Networking Conference (JENC8) Edinburgh Scotland in May, 1997 (Optical StarTAP)
  - **Optical Nets:** State of art DWDM over fiber for massive bandwidth
  - **Lambda Grids / Lambda Fabrics:** Circuit switching to provide end-to-end paths for distributed services *(now production with ESNet OSCARS w/VLANs)*
- ***Now*: Resource Disaggregation and Serverless computing**
  - Seeing lambda grid concept emerging within rack & chip
  - Miniaturized DWDM now within a 5x7mm silicon die! *(smaller than a dime)*
  - Optical Circuit Switching and Lambda-steering within chip and rack

Diverse Node Configurations for Diverse Workload Resource Requirements

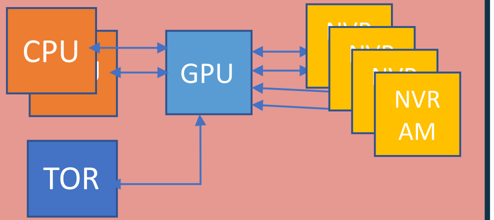
Training

- 8 connections: GPU
- 8 links to HBM (weights)
- 8 links: to NVRAM
- 1 links: to CPU (control)



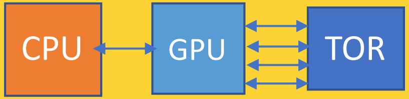
Data Mining

- 6-links: HBM
- 15 links: NVRAM (capacity)
- 4 links: CPU (branchy code)



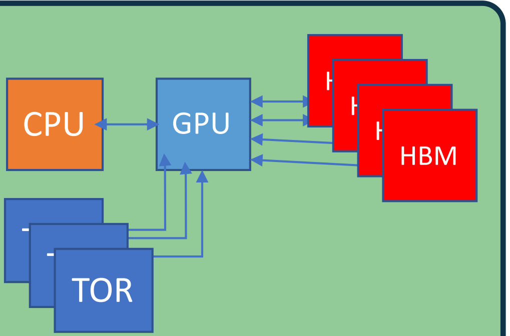
Inference

- 16 links to TOR (streaming data)
- 8 links HBM (weights)
- 1 link: CPU



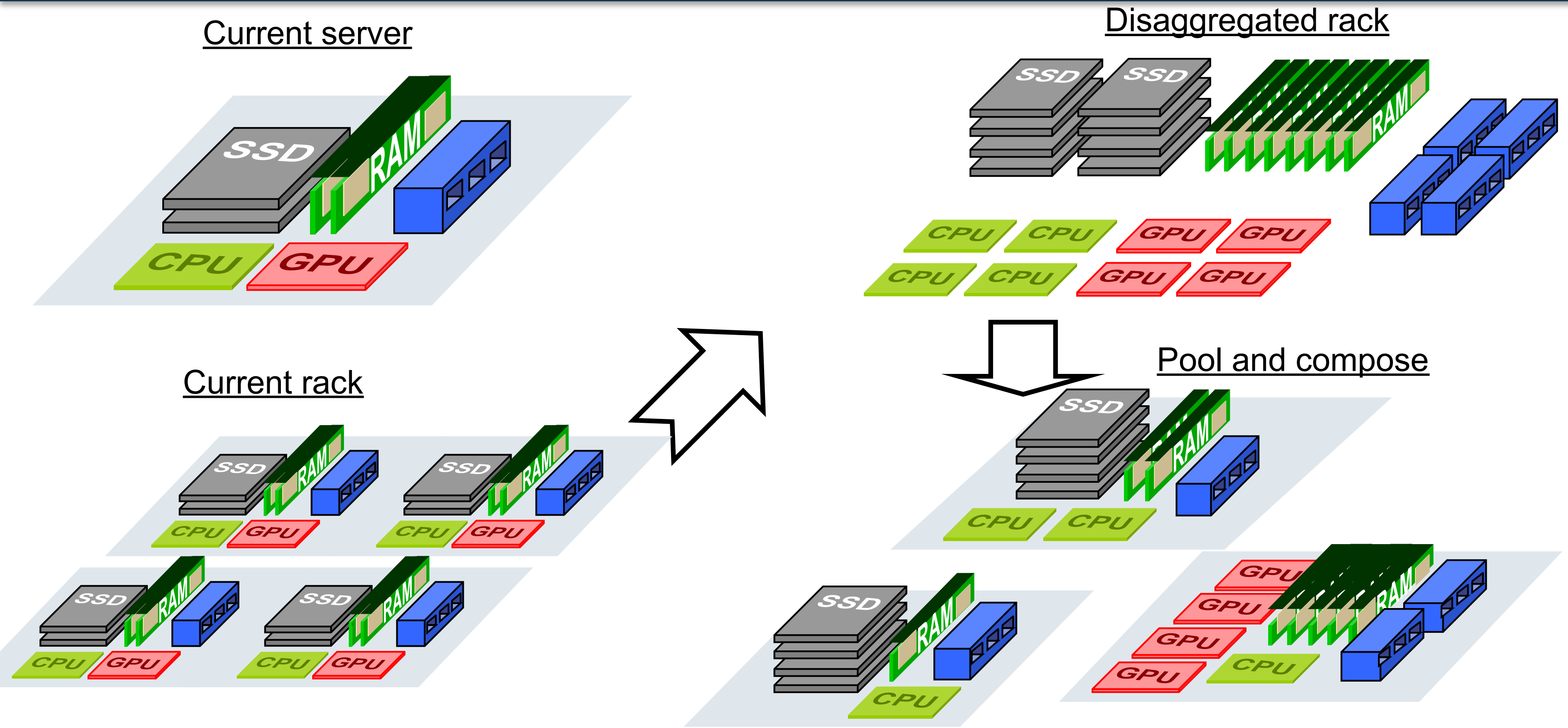
Graph Analytics

- 16 links HBM
- 8 links TOR
- 1 Link CPU



TOR NVRAM GPU HBM CPU

Disaggregated Node/Rack Architecture



Current server

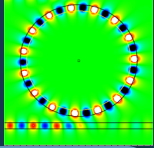
Current rack

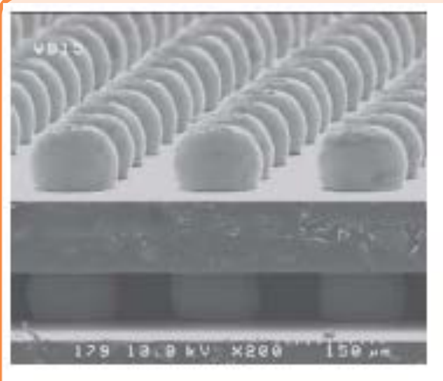
Disaggregated rack

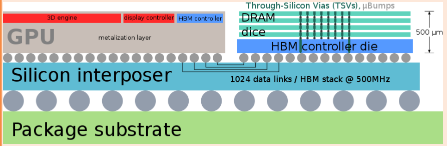
Pool and compose

Most solutions current disaggregation solutions use Interconnect bandwidth (1 – 10 GB/s)
But this is significantly inferior to RAM bandwidth (100 GB/s – 1 TB/s)

DWDM has moved inside of the chip!



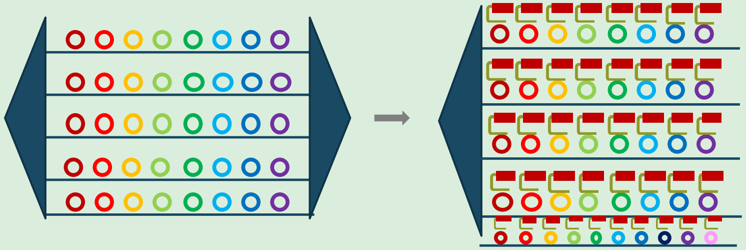


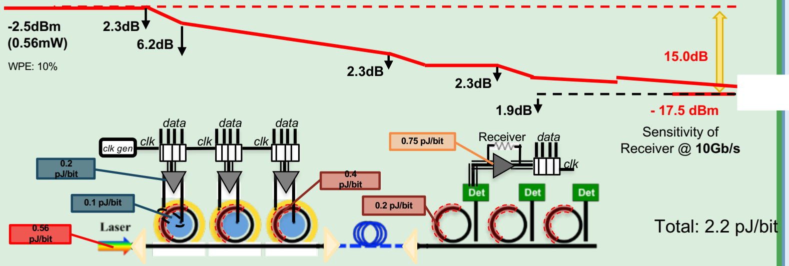


In-package integration

Solder Microbumps
& Copper Pillars@~10Gb

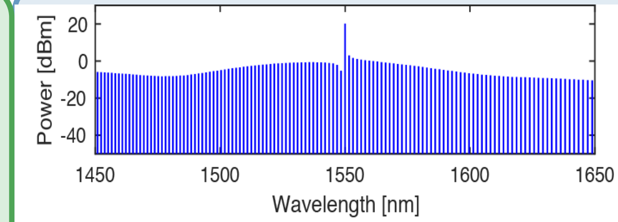
Wide and Slow!

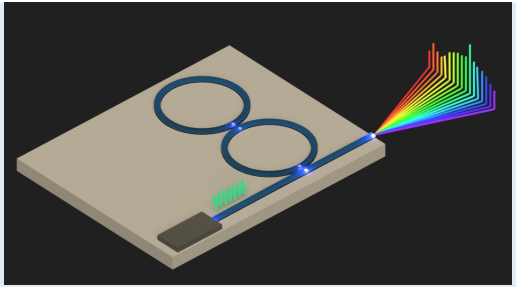




DWDM Using Silicon Photonics

Ring Resonators @ ~10-25 Gb/sec per chan





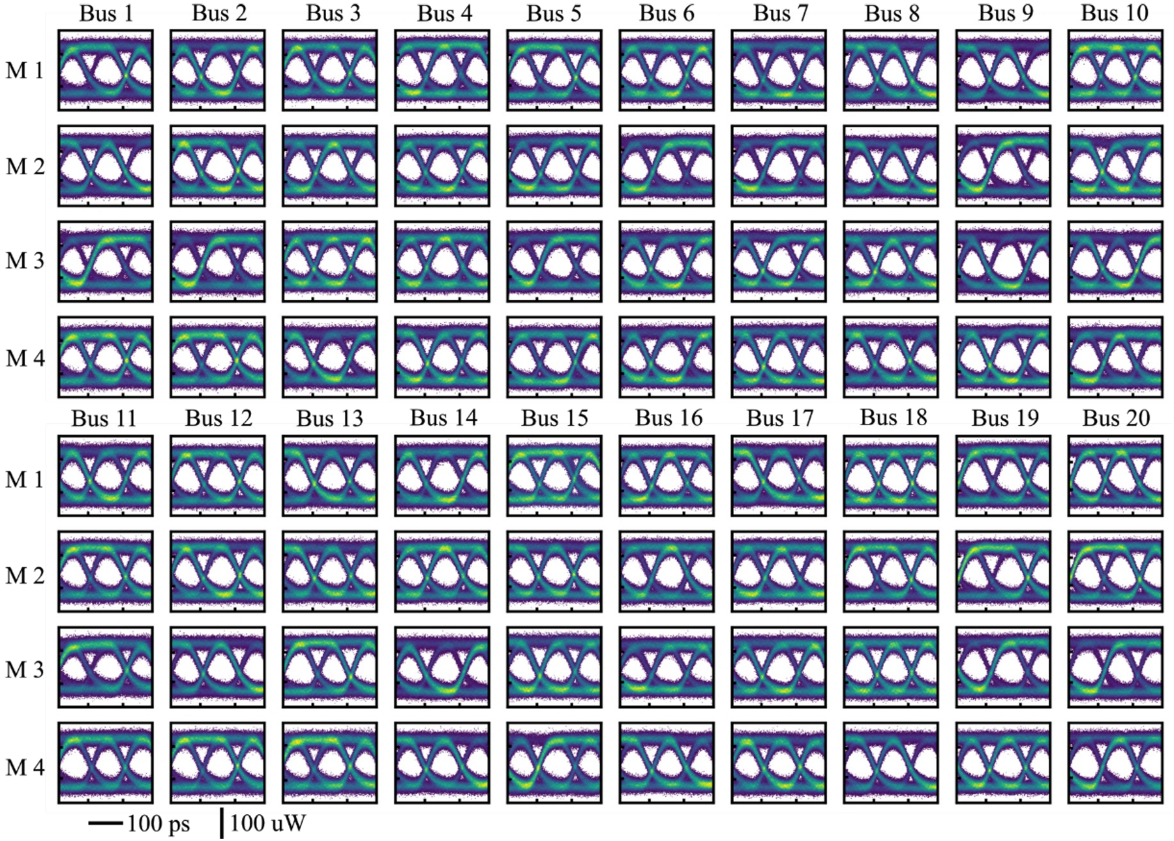
Comb Laser Sources

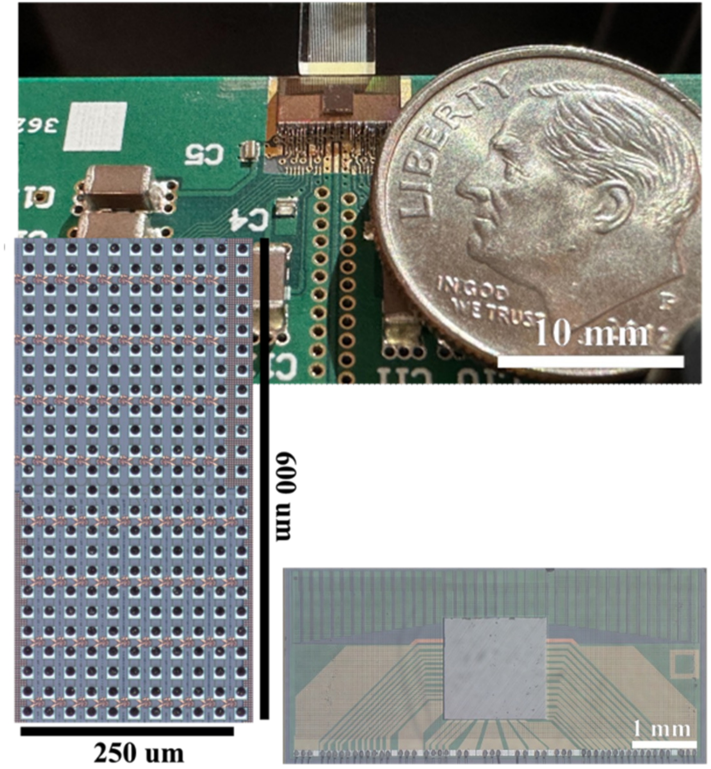
Single laser to efficiently
nerate 100s of frequencies

de and Slow!

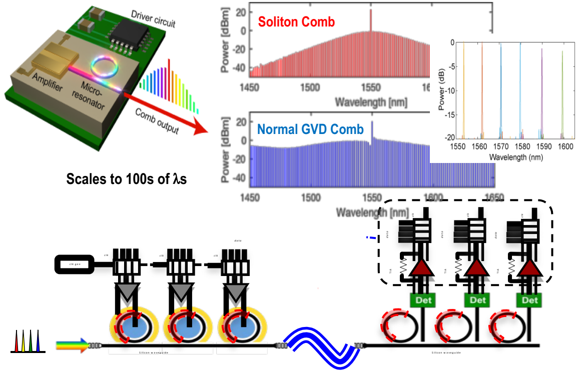
And like with StarTap and TransLight and other DWDM Lambda-grids, that kind of BW it opens up so many new possibilities!

> 5Tb/mm^2 Bandwidth Density 3D 80-channel Transmitter



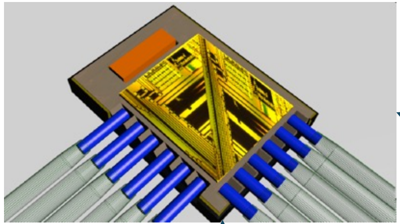


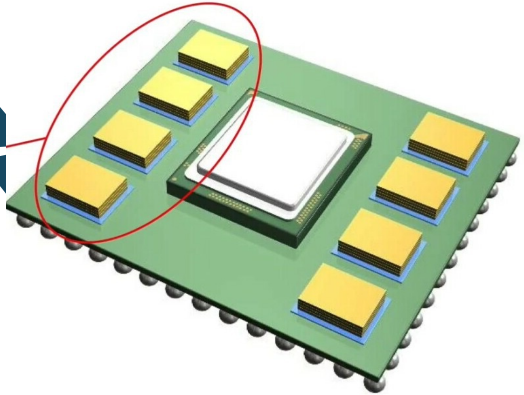
Photonic MCM (Multi-Chip Module)



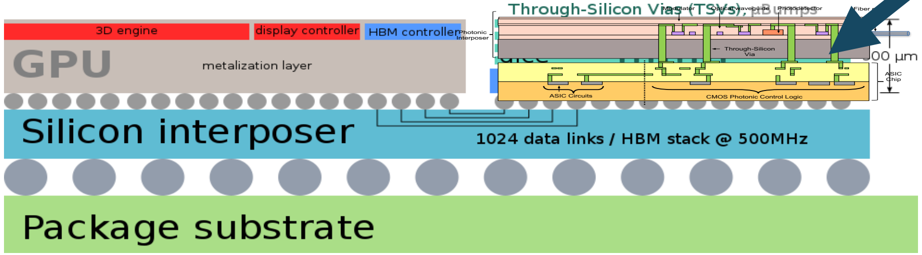
Comb Laser Source with DWDM Silicon Photonics

Wide-and Slow for high speed links





Photonic SiP



3D engine

display controller

HBM controller

GPU

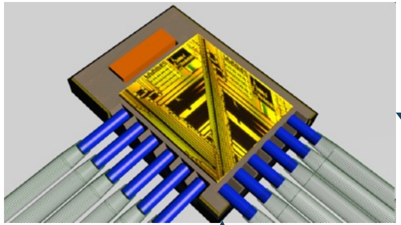
metalization layer

Silicon interposer

1024 data links / HBM stack @ 500MHz

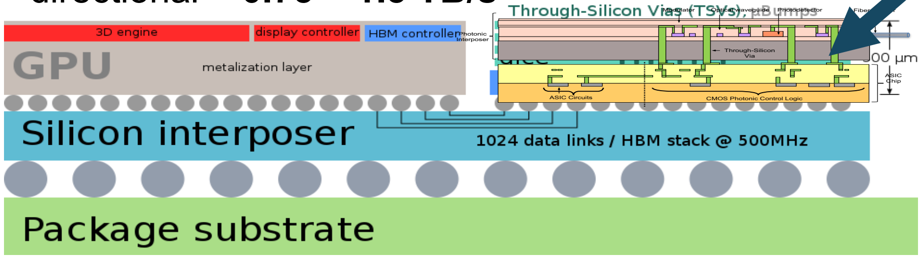
Package substrate

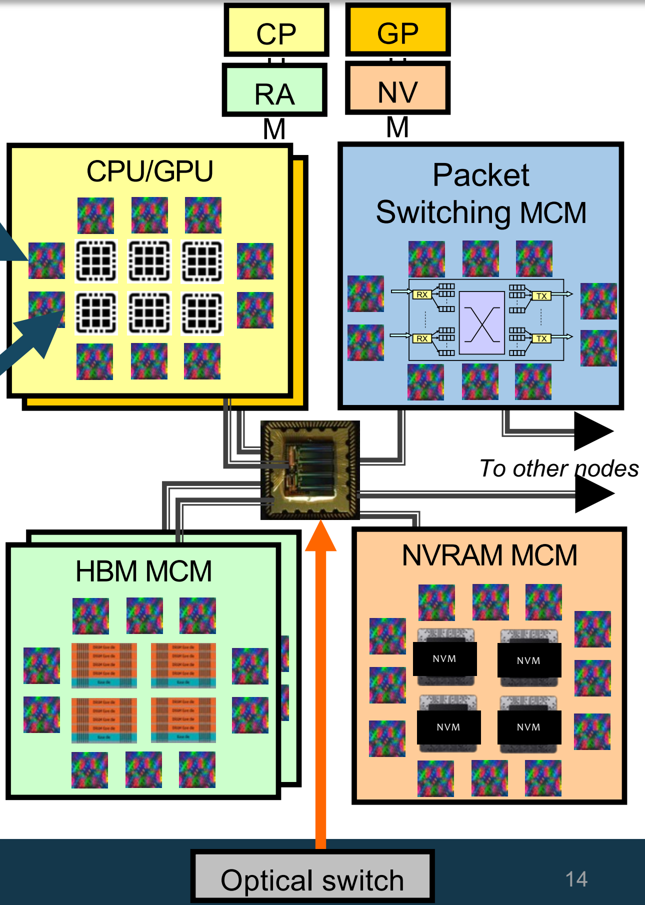
Photonic MCM (Multi-Chip Module)



High-Density **fiber coupling array** with 24 fibers = 6-12 Tb/s bi-directional = **0.75 – 1.5 TB/s**

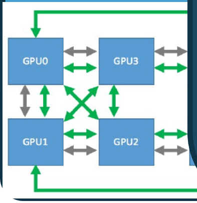
Photonic SiP





T

- 8 conn
- 8 links
- 8 links
- 1 links



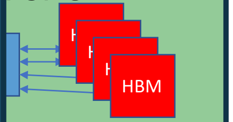
Emerging disaggregated datacenter architectures

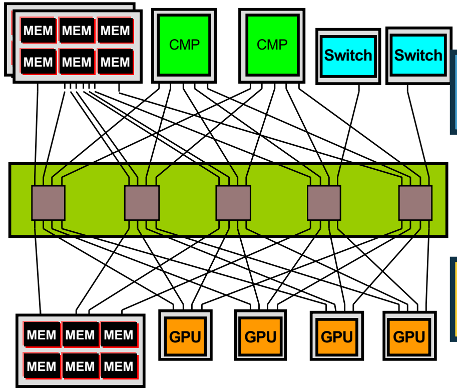
- Its all about the data flow!
- Revisit network description languages for optical networks
- Role based control models for multi-domain apps
- Scalable workflows

- And Security!!!!

Analytics

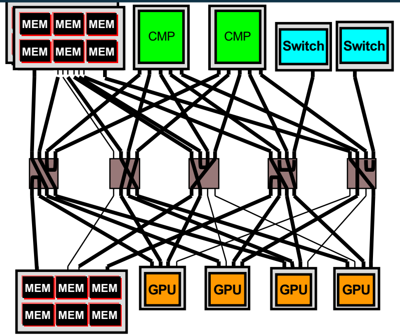
ks HBM
s TOR
k CPU

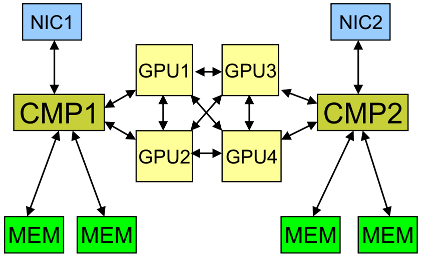


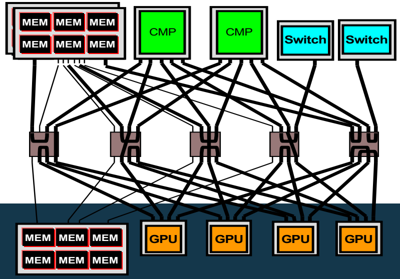


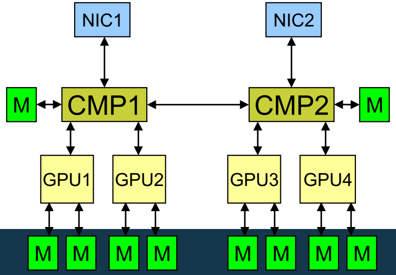
Configure for Training

Configure for Inference









Other Consequences of Disaggregation

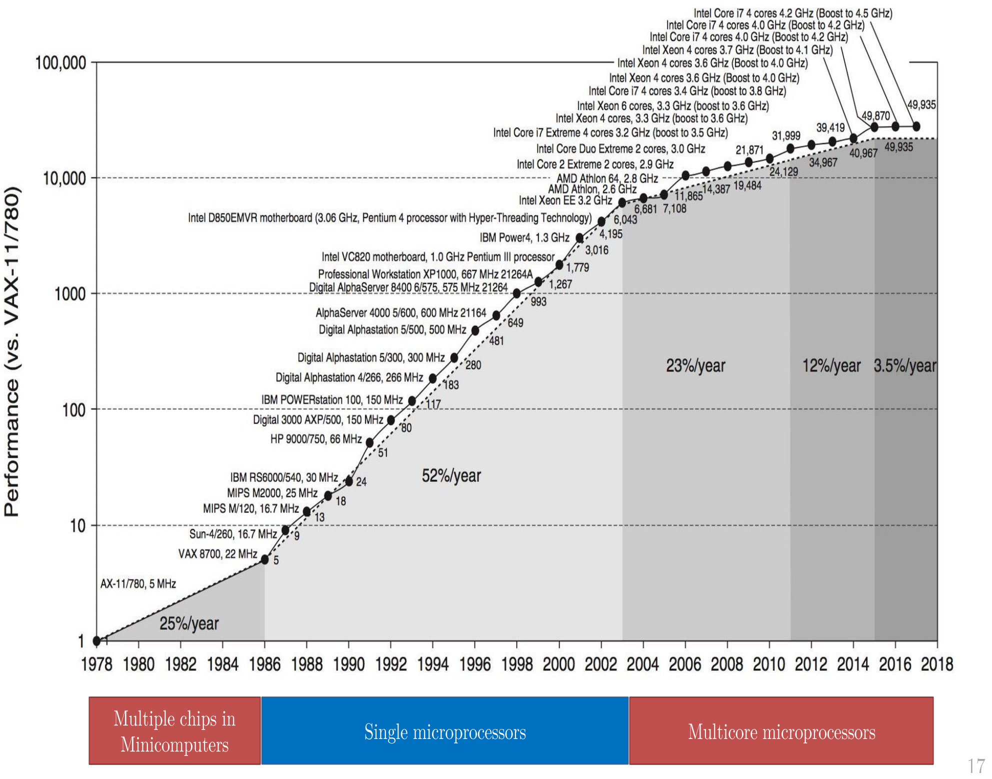
- **Security**
  - **Conventional Wisdom: B**oundaries of Linux server are the DMZ
  - **New World Order:** what boundary? What Linux Server?

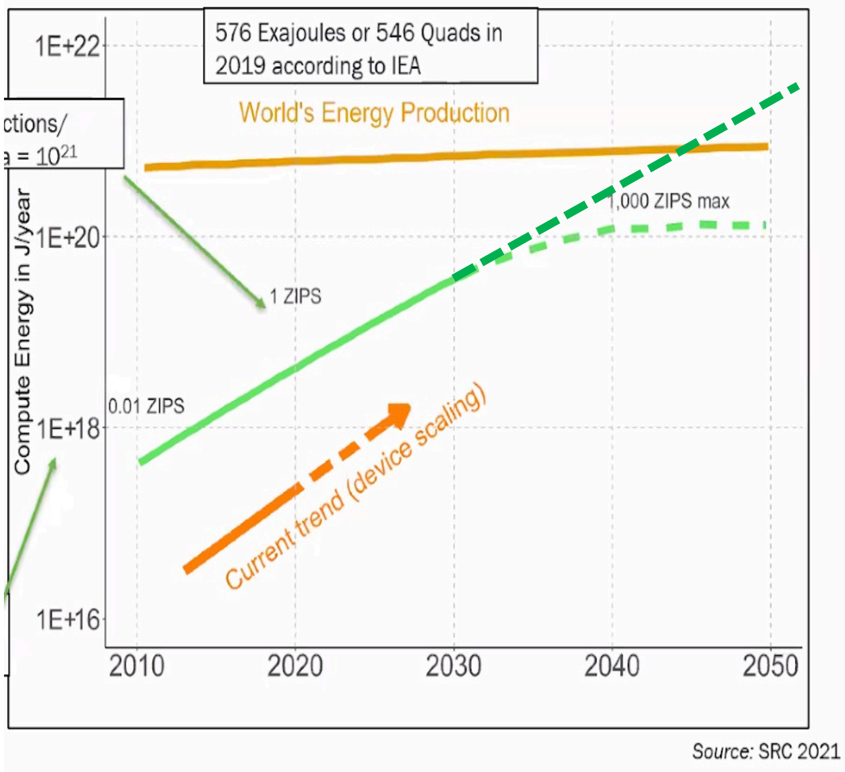
- **Emerging Trusted Execution Environments (TEEs)**
  - *Now all resources are distributed and must have a ”shared secret” to work together safely*
  - *Trust-no-one… revocable credentials, differential security*
  - *Solutions to security even within the rack of this new “disaggregated datacenter” are looking like the iWAY and Grid and modern wide area distributed services*

- **Emerging Technology looks a LOT like “déjà vu All Over Again**

Moore’s Law is Ending *(really it is!)*

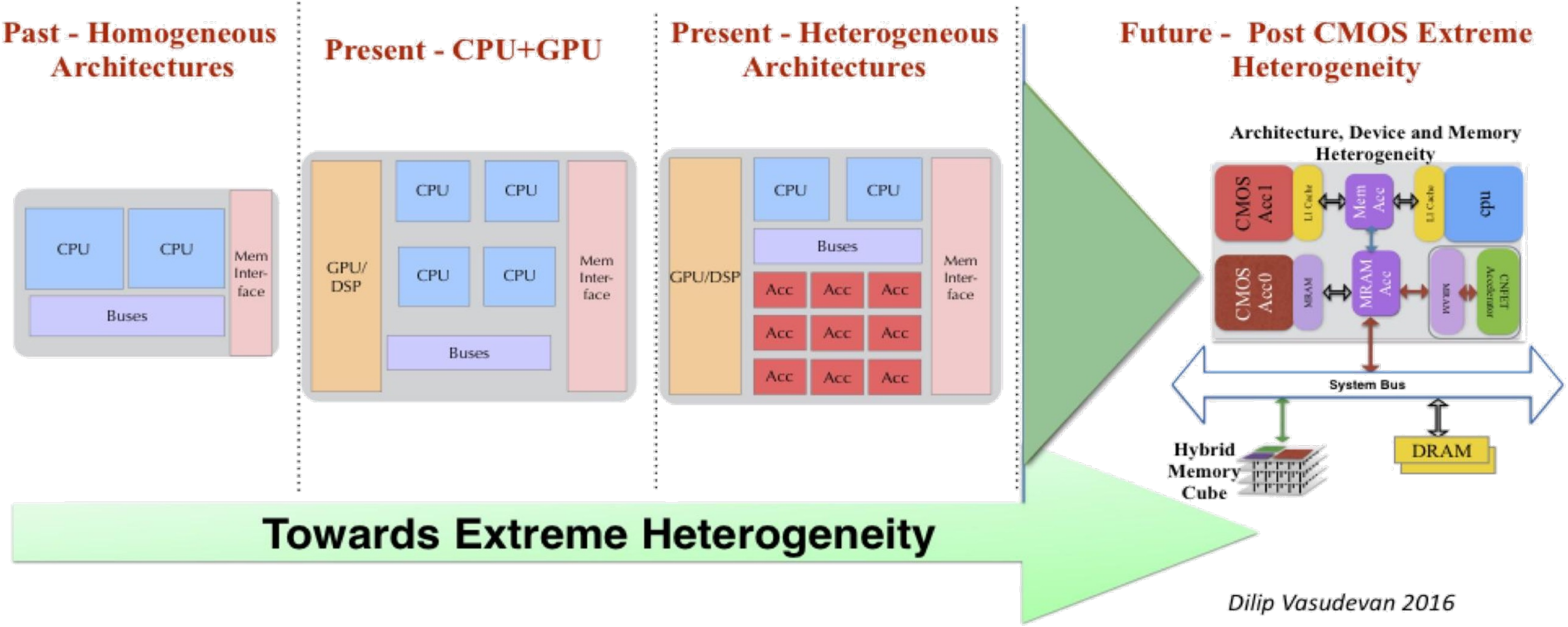
Hennessy / Patterson





Source: SRC 2021

The Future Direction for Post-Exascale Computing



Dilip Vasudevan 2016

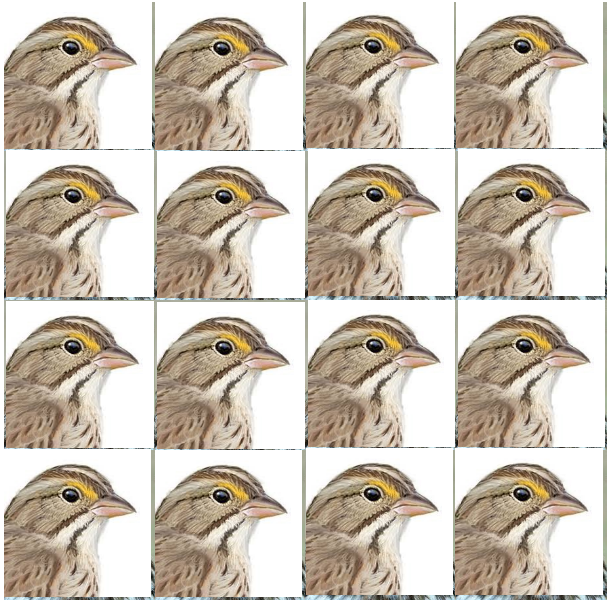
Specialization:

Natures way of Extracting More Performance in Resource Limited Environment

Powerful General Purpose

Xeon, Power

Many Lighter Weight
(post-Dennard scarcity)



KNL, AMD, Cavium/Marvell, GPU

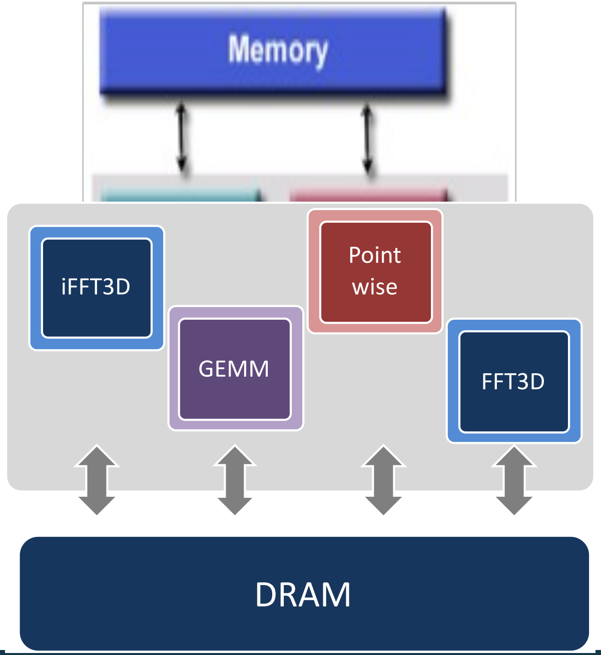
Many Different Specialized
(Post-Moore Scarcity)

Probing
Filter feeding
Generalist
Insect catching
Grain eating
Coniferous-seed eating
Aerial fishing
Pursuit fishing
Nectar feeding
Fruit eating
Chiseling
Dip netting
Scavenging
Raptorial
Surface skimming
Scything

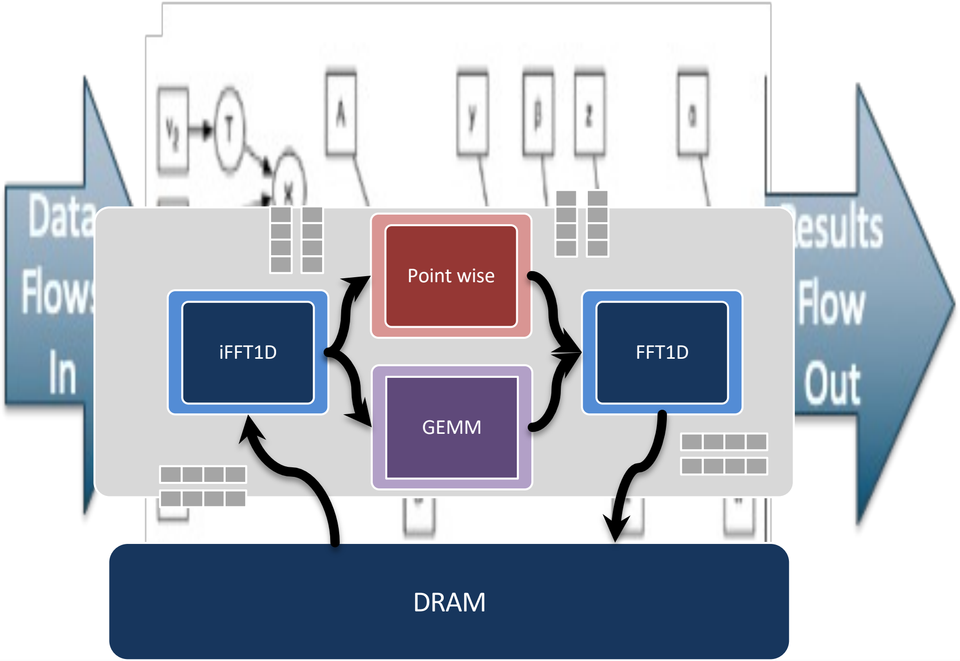
Apple, Google, Amazon

Algorithm Reformulated as Custom Circuit

Von Neumann CPU



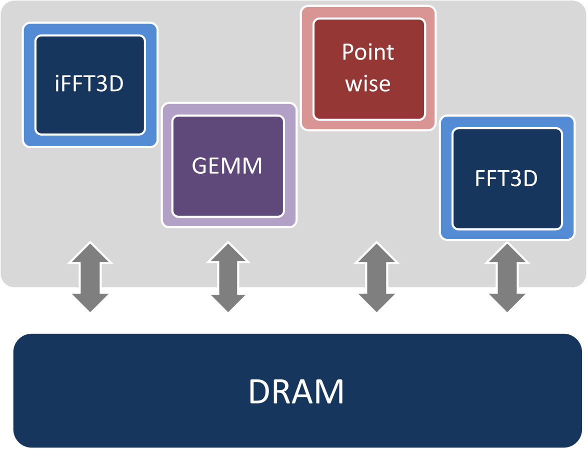
Dataflow (FPGA, GraphCore etc.)

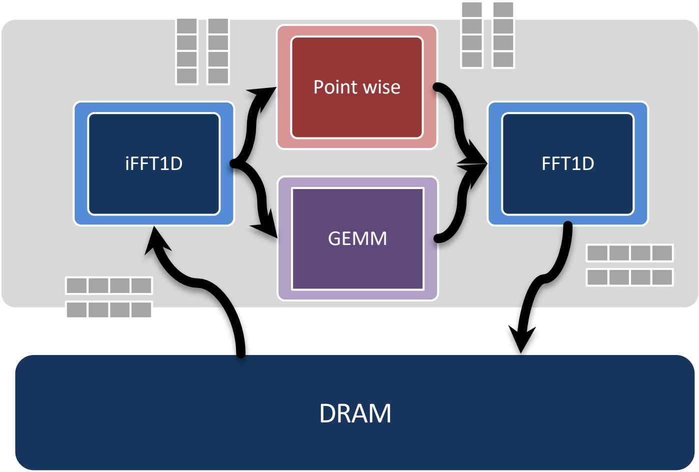


See Also Torsten Hoefler “StreamBLAS” for FPGA

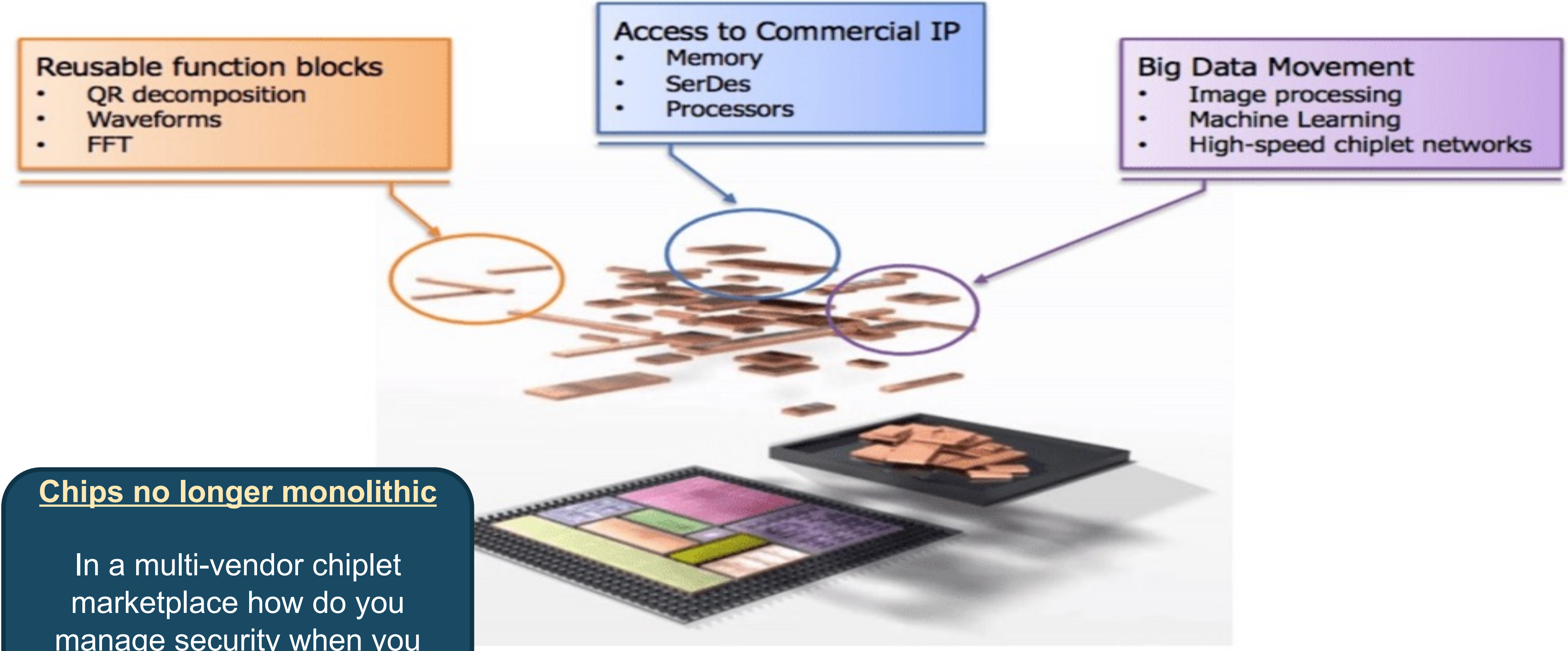
From GSDL to WSDL to Workflows

- **Doesn't this look kind of familiar?**
  - Moving SaaS, FaaS, and *aaS towards workflows
  - Wide area networking has at least 2 decades lead thinking through these complex issues of service orchestration!





How do chiplets enable domain specialization?



Reusable function blocks
- QR decomposition
- Waveforms
- FFT

Access to Commercial IP
- Memory
- SerDes
- Processors

Big Data Movement
- Image processing
- Machine Learning
- High-speed chiplet networks

Chips no longer monolithic

In a multi-vendor chiplet marketplace how do you manage security when you can’t trust all of your chiplets ?

CHIPS modularity targets the enabling of a wide range of custom solutions

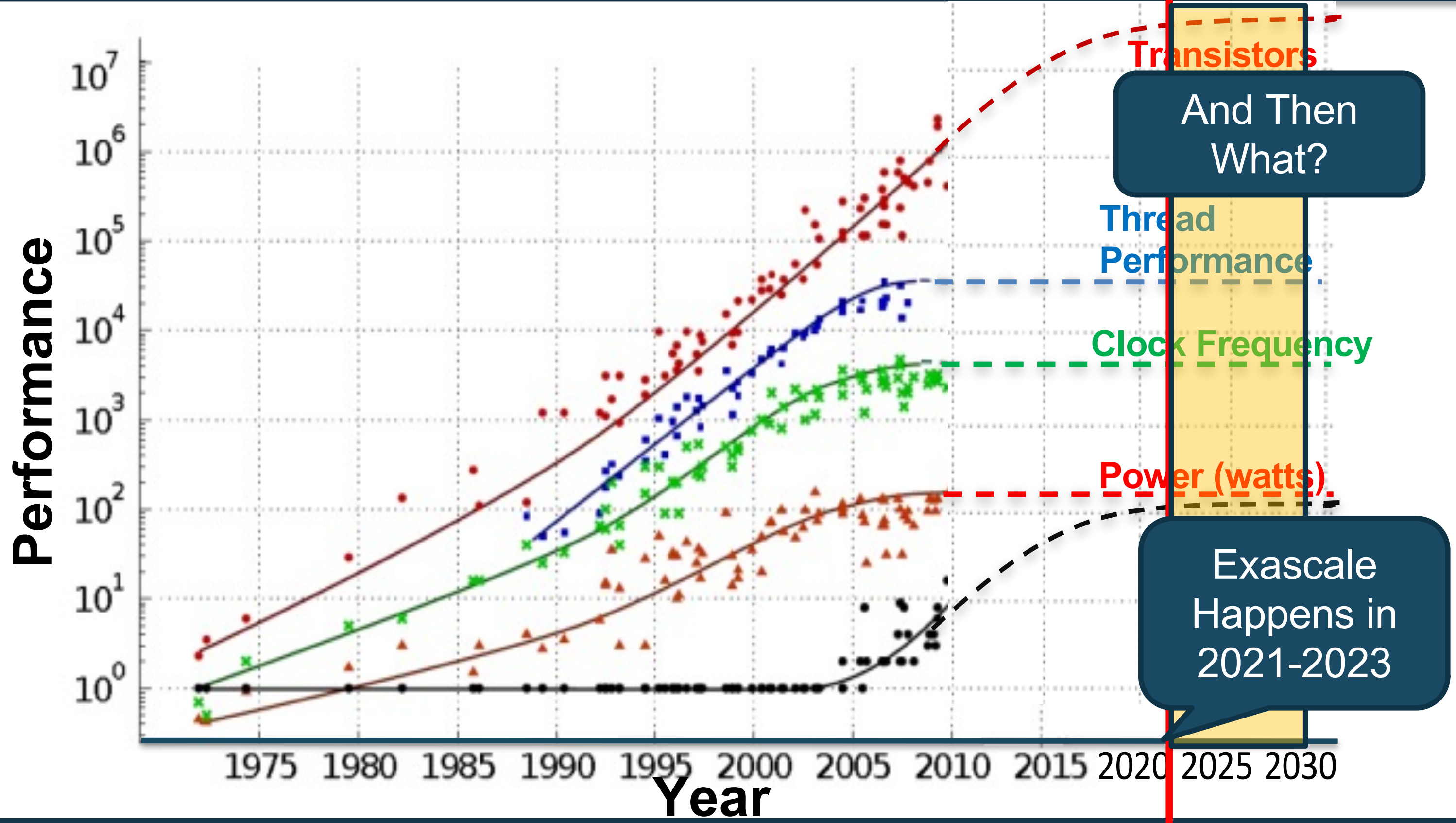
The future looks more like the past

- **The slowdown in Moores Law isdriving a new world order in datacenters!**
  - Disaggregation, extreme heterogeneity, serverless computing, break-down of security models
- **Wide Area High Performance optical networks and Distributed Services architectures have had to grapple with these issues for decades before**
  - Lambda-switching/steering
  - Workflow description and service orchestration
  - Distributed “trust no-one” **security** and differential privacy models (inside chip!!!)
  - *as –a-Service models (Accelerator as a Service for example)
- ***Cees and Leon could easily dominate next generation of computer architecture research just by drawing on their ample (30+ years) of accumulated knowledge of wide area distributed computing…. (another 30+ years of work ahead)***

Its like déjà vu all over again –*Yogi Berra*

Technology Scaling Trends

Exascale in 2021… and then what?

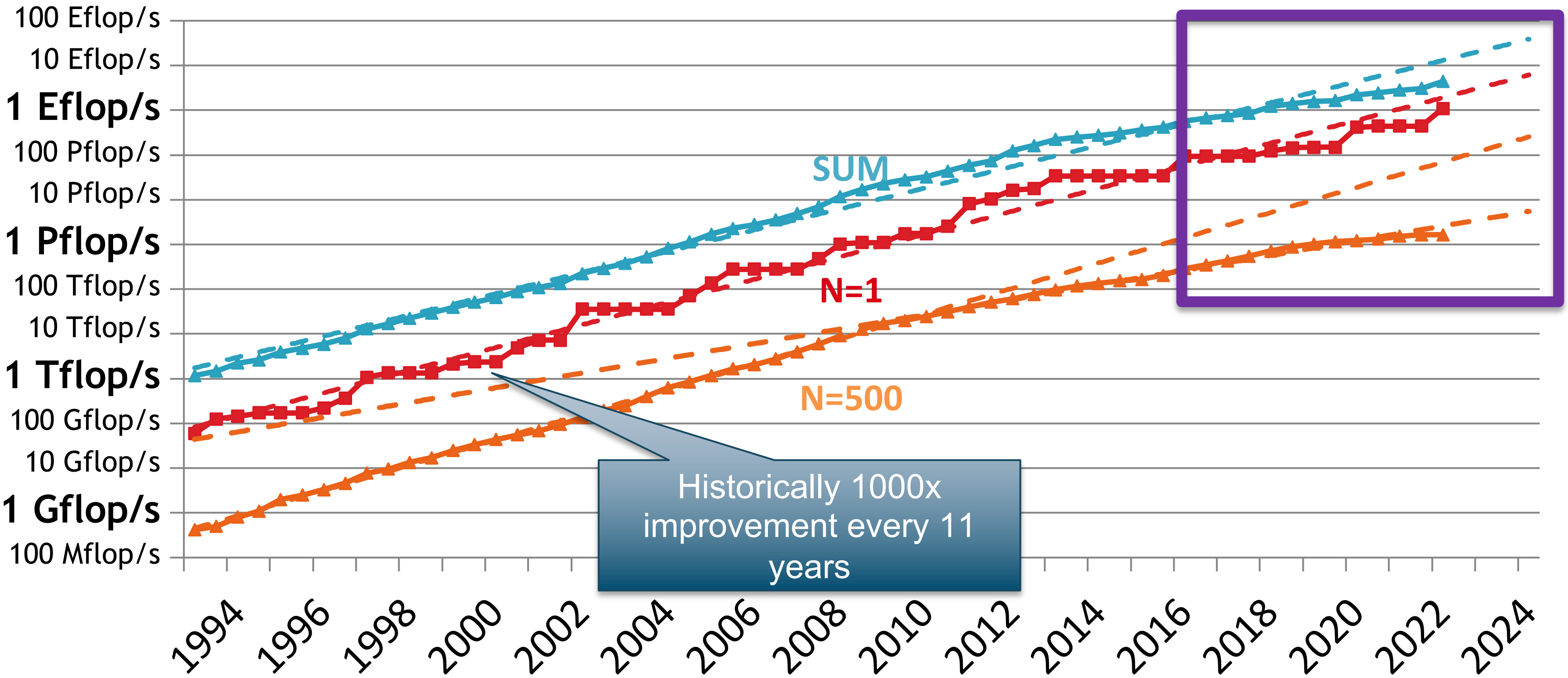


And Then What?

Exascale Happens in 2021-2023

Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

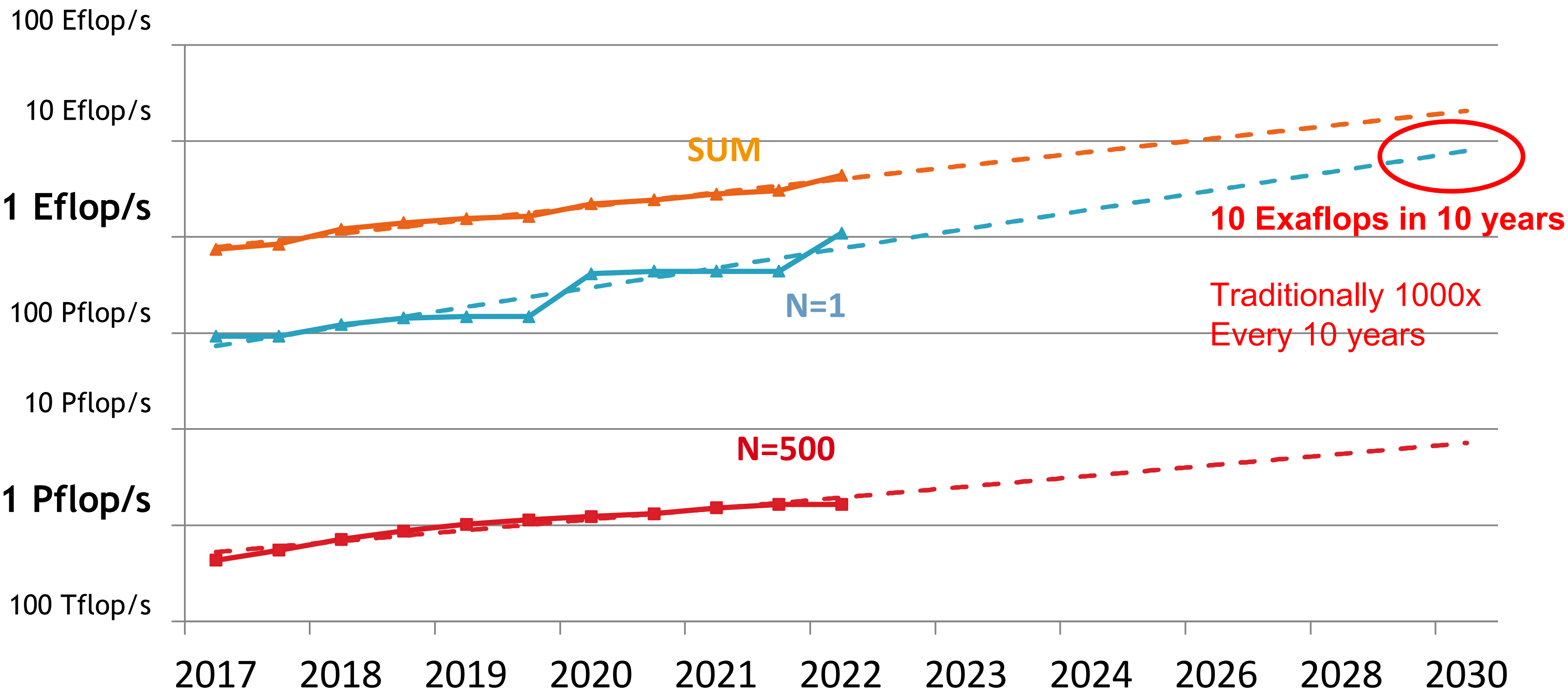
Projected Performance Development



Historically 1000x improvement every 11 years

Erich Strohmaier
Top500.org

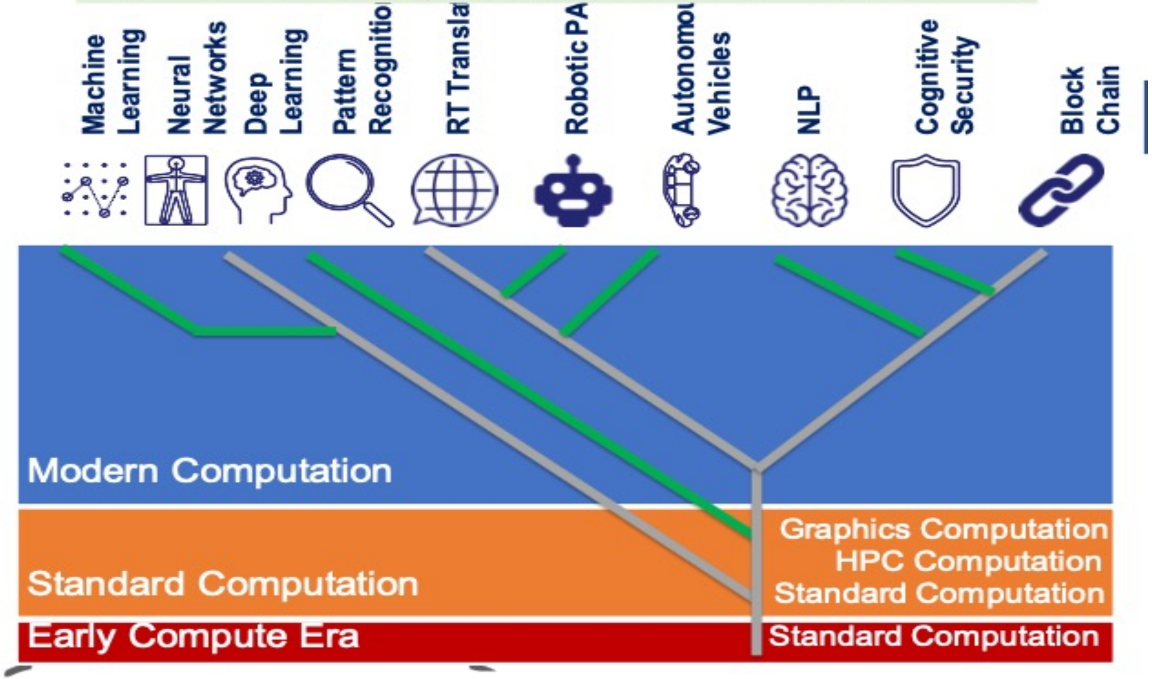
Projected Performance Development



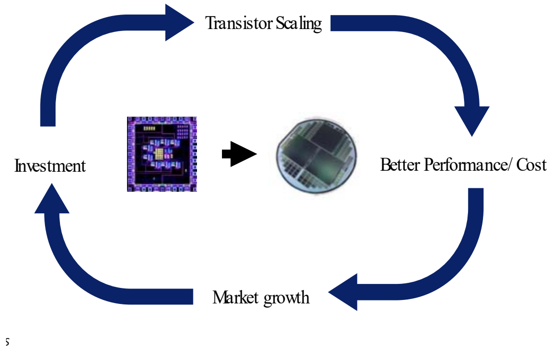
Why? Domain specific Architectures driven by hyperscalers

in response to slowing of Moore's Law (switch to systems focus for future scaling)

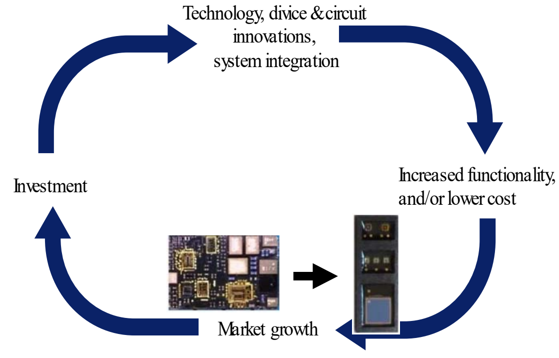
Dharmesh Jani, Facebook –
ODSA Workshop, Regional Summit, Amsterdam, Sep. 2019



AI/ML/data workload explosion needs DSAs

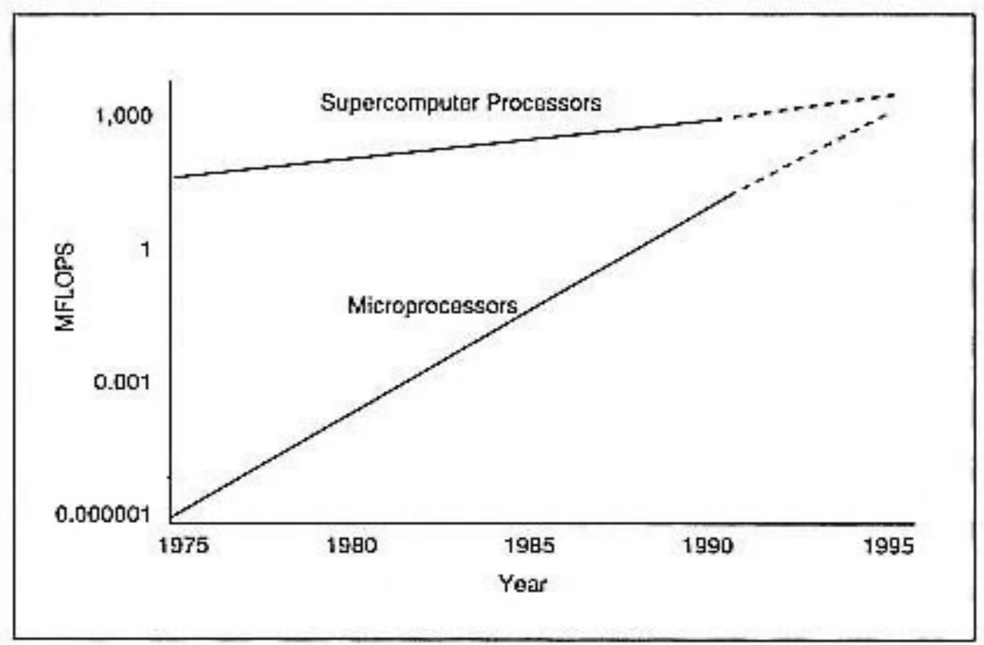


Transistor Focus


Technology, divice & circuit innovations,
system integration
Increased functionality, and/or lower cost
Market growth
Investment

System Focus

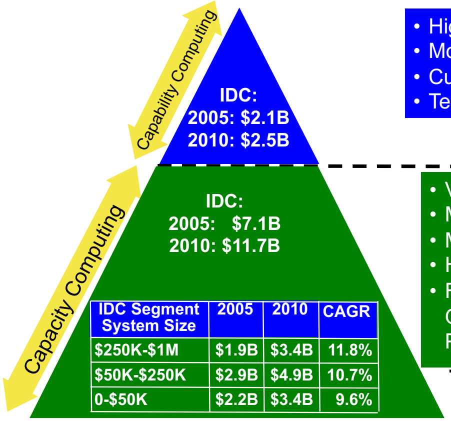
Attack of the Killer Micros



Attack of the killer micros

John Markoff, May 6, 1991

- Was more about the economic model than technology alone



IDC:
2005: $2.1B
2010: $2.5B

IDC:
2005: $7.1B
2010: $11.7B

IDC Segment System Size	2005	2010	CAGR
$250K-$1M	$1.9B	$3.4B	11.8%
$50K-$250K	$2.9B	$4.9B	10.7%
0-$50K	$2.2B	$3.4B	9.6%

- High End Systems (>$1M)
- Most/all Top 500 systems
- Custom SW & ISV apps
- Technology risk takers & early adopte

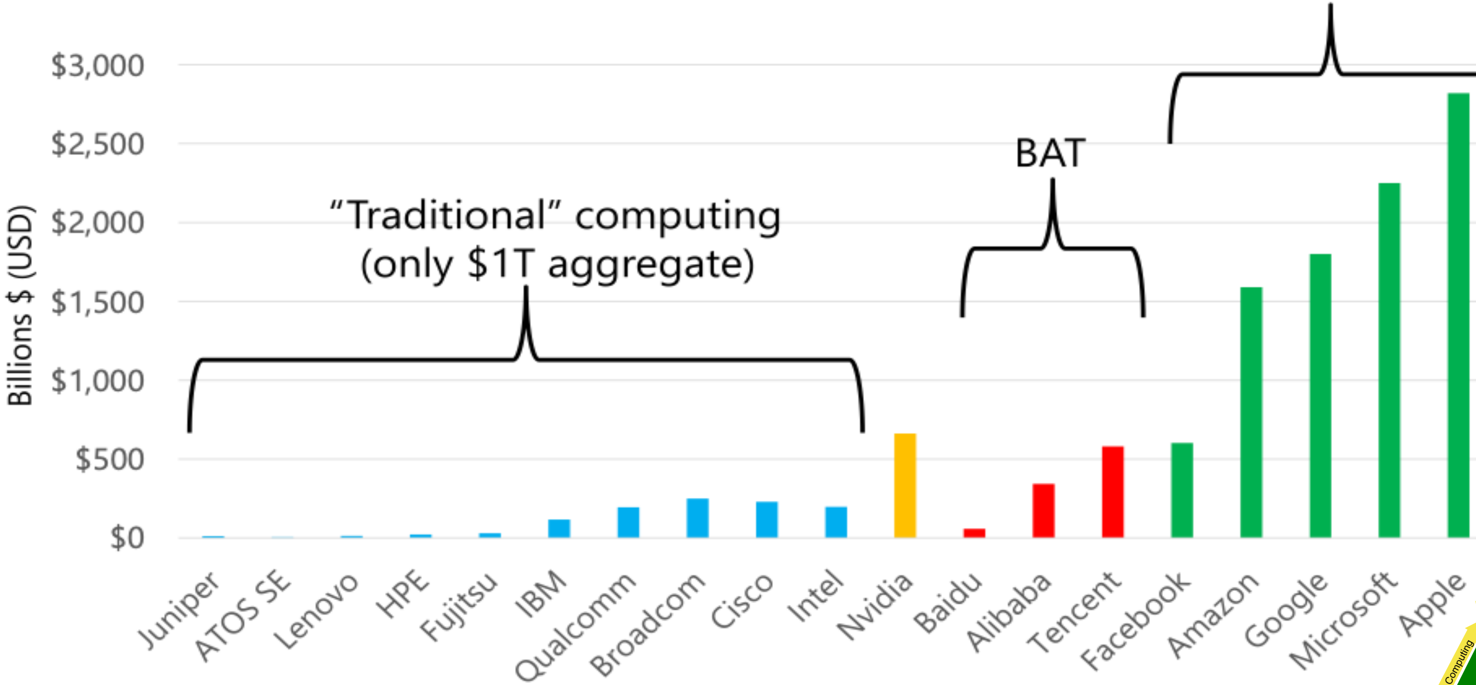
- Volume Market
- Mainly capacity; <~150 nodes
- Mostly clusters; >50% & growing
- Higher % of ISV apps
- Fast growth from commercial HPC; Oil &Gas, Financial services, Pharma, Aerospace, etc.

Total market >$10.0B in 2006
Forecast >$15.5B in 2011

HPC is built with of pyramid investment model

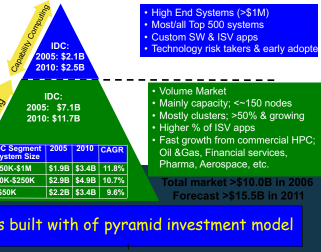
It is not good enough anymore to understand the technology Now we must also understand the market context

Dan Reed, 2022
https://arxiv.org/pdf/2203.02544.pdf



Control of the computing ecosystem

Trillion+ $ (USD) companies



IDC Segment System Size	2005	2010	CAGR
$250K-$1M	$1.9B	$3.4B	11.8%
$50K-$250K	$2.9B	$4.9B	10.7%
0-$50K	$2.2B	$3.4B	9.6%

- High End Systems (>$1M)
- Most/all Top 500 systems
- Custom SW & ISV apps
- Technology risk takers & early adopte

- Volume Market
- Mainly capacity; <~150 nodes
- Mostly clusters; >50% & growing
- Higher % of ISV apps
- Fast growth from commercial HPC; Oil &Gas, Financial services, Pharma, Aerospace, etc.

Total market >$10.0B in 2006
Forecast >$15.5B in 2011

HPC is built with of pyramid investment model

Opportunity for HPC: New Economic Model

Open Chiplets Marketplace is forming (ODSA and UCIexpress)

- Licensable IP and assembly by 3rd party lowers that barrier
- Leverage the economic model being created by HyperScale

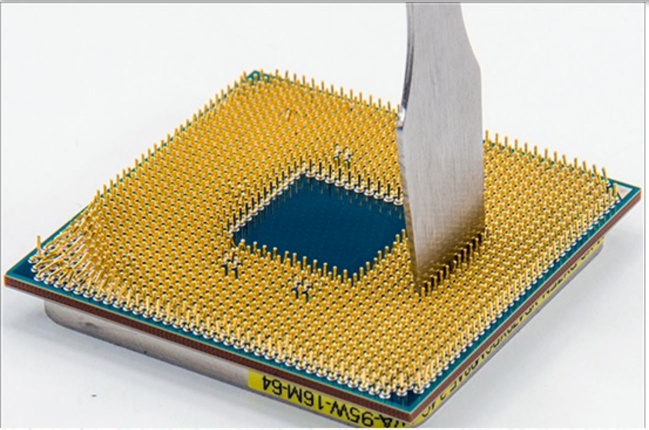
Leverage this baseline and extend to support HPC

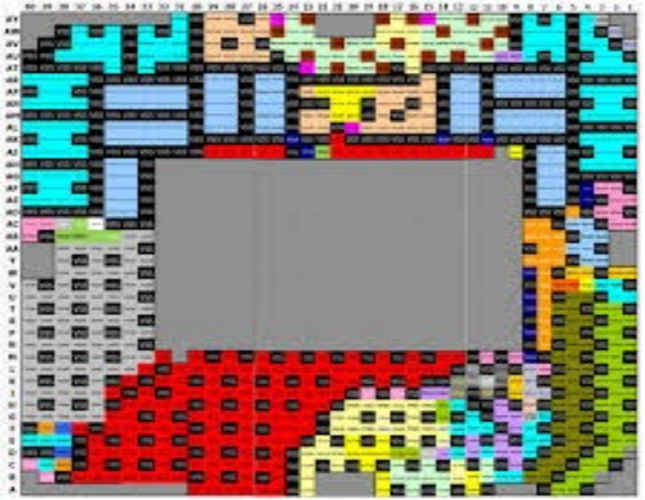
- Smaller incremental cost for HPC to "play"
- *HPC has become "too small to attack the city"*

80:20 Rule: Focus open efforts on what uniquely benefits HPC

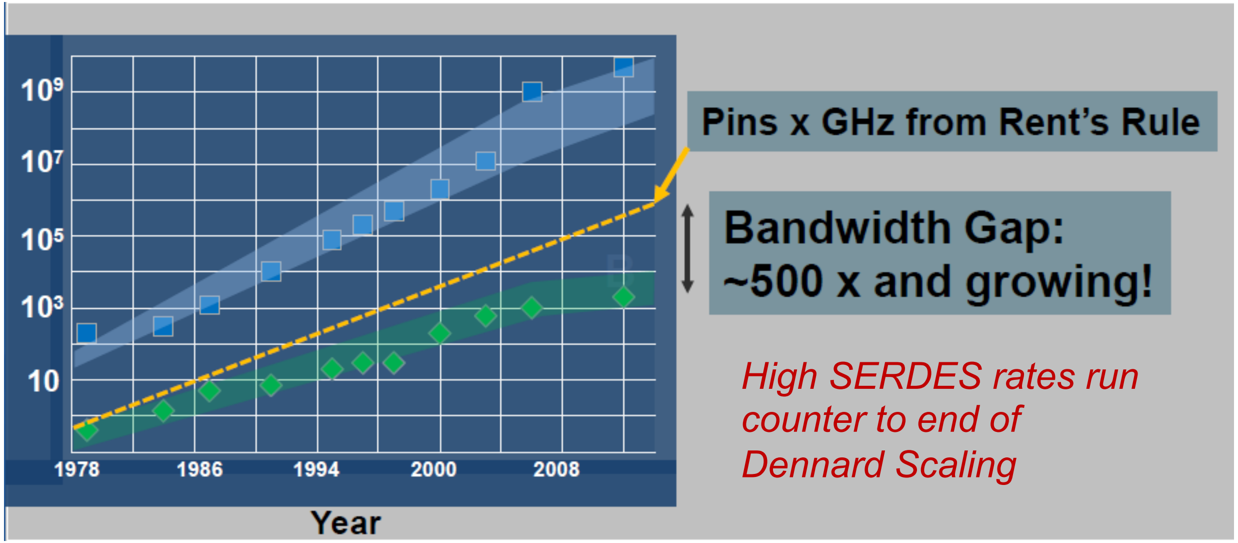
- Build up a library of reusable accelerators for HPC.
- Interoperability for sustainability: *Interoperate with Arm IP for commercially supported IP where it exists and focus Open on the 20% that doesn't make commercial sense to license*

Package Performance is Pin Limited





Rent's Rule:
Number of pins = K x Gates^a (IBM, 1960)
K = 0.82, a = 0.45 for early Microprocessors



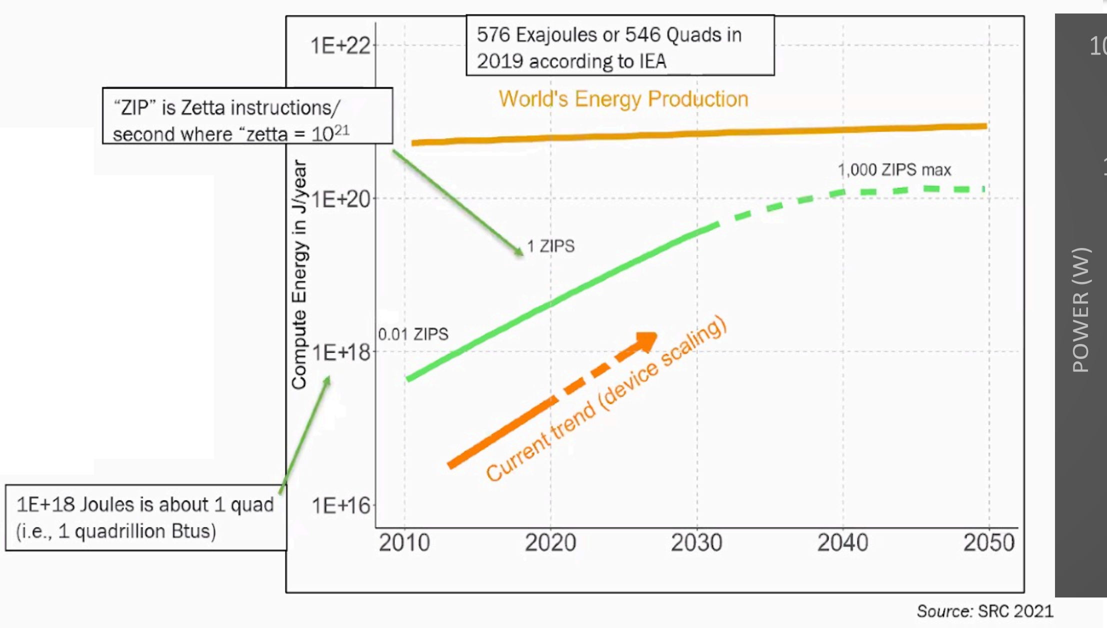
Pins x GHz from Rent's Rule

Bandwidth Gap:
~500 x and growing!

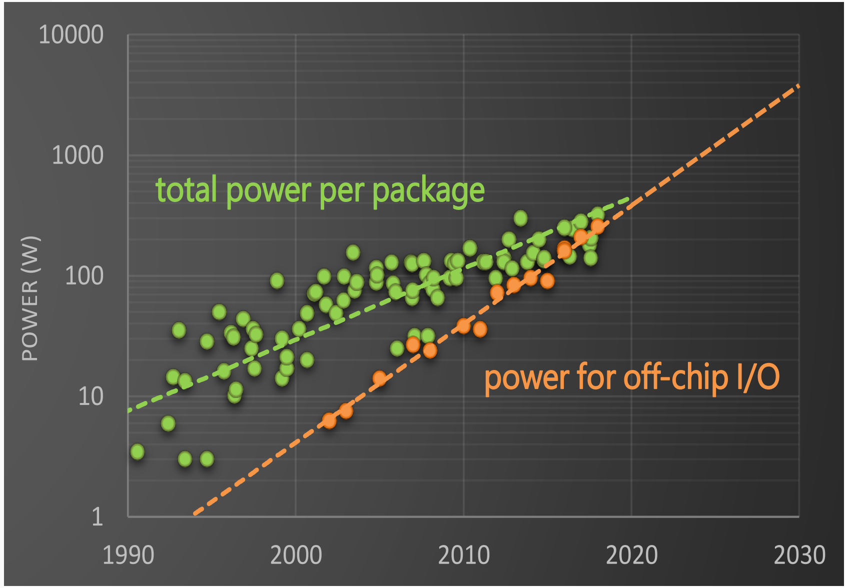
High SERDES rates run counter to end of Dennard Scaling

Source: J. Poulton, Nvidia

Datacenters: Worsten climate change without ultra-energy-efficiency And data movement dominates that power consumption



Source: SRC 2021



Source: Gordon Keeler (DARPA)

- **January 2021 SRC report projects datacenter energy growth rates will lead to ~25% consumption of planetary energy by 2040.**
- **Data movement is a dominant contributor to that power consumption**