Heterogeneous Integration for HPC and Data Centers

TWG Chair: Kanad Ghose, Ph.D.

Distinguished Professor of Computer Science, SUNY-Binghamton
PhD (CS), M.Tech (EE)
and Site Director of Center for Energy-Smart Electronic Systems,
a NSF Industry/University Collaborative Research Center

TWG Co-Chair: John Shalf.

Department Head for Computer Science and Computer Engineering
Lawrence Berkeley National Laboratory

MS (EE/CE)

http://eps.ieee.org/technology/heterogeneous-integration-roadmap.html

eps.ieee.org/hir-2021

Intent of the Chapter and Notes

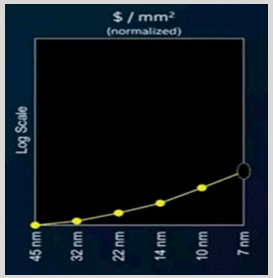
- **Heterogeneous integration is not just about coming up with a packaging solution to house connected chiplets**

  Its all about systems architecture and integration

- There are **many crosscutting issues** that need to be considered as part of the packaging solution:
  - Diversity of chiplets
  - Interconnections
  - Power conversion and delivery
  - Security issues
  - Other considerations, including QC systems
  - Verification/test, Design automation
- **Not possible to come up with generations, quantification of trends of all factors**, but expected trends can be shown for some
  - System architectures/ components evolving continuously
  - Tying trends to a timeline is difficult

HI Drivers in the HPC/Data Center Market

Increasingly higher cost per unit area of large dies



Application-Specific Accelerators and New Memory Technologies

Lack of expected performance and energy efficiency scaling with node advances

HETEROGENEOUSLY INTEGRATED PRODUCTS FOR THE HPC/DATA CENTER SEGMENT

Increasing Reliance on the Web and Emerging Applications:

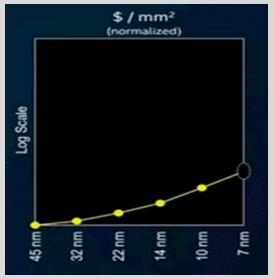
- Analytics/Intelligence on demand
- Big data processing
- IoTs and Edge
- Blockchain processing

Increasing need to go green, energy supply constraints

IEEE | IEEE photonics Society | semi | IEEE ELECTRONICS PACKAGING SOCIETY | ASME | ELECTRON DEVICES SOCIETY

Emphasis area for planned 2024 Updates

Increasingly higher cost per unit area of large dies



Application-Specific Accelerators and New Memory Technologies

Lack of expected performance and energy efficiency scaling with node advances

HETEROGENEOUSLY INTEGRATED PRODUCTS FOR THE HPC/DATA CENTER SEGMENT

Increasing Reliance on the Web and Emerging Applications:

- **Analytics/Intelligence on demand**
- Big data processing
- IoTs and Edge
- Blockchain processing

Energy supply constraints, Sustainability

IEEE IEEE photonics Society semi IEEE ELECTRONICS PACKAGING SOCIETY ASME ELECTRON DEVICES SOCIETY

Data Center, AI Energy Consumption Trends

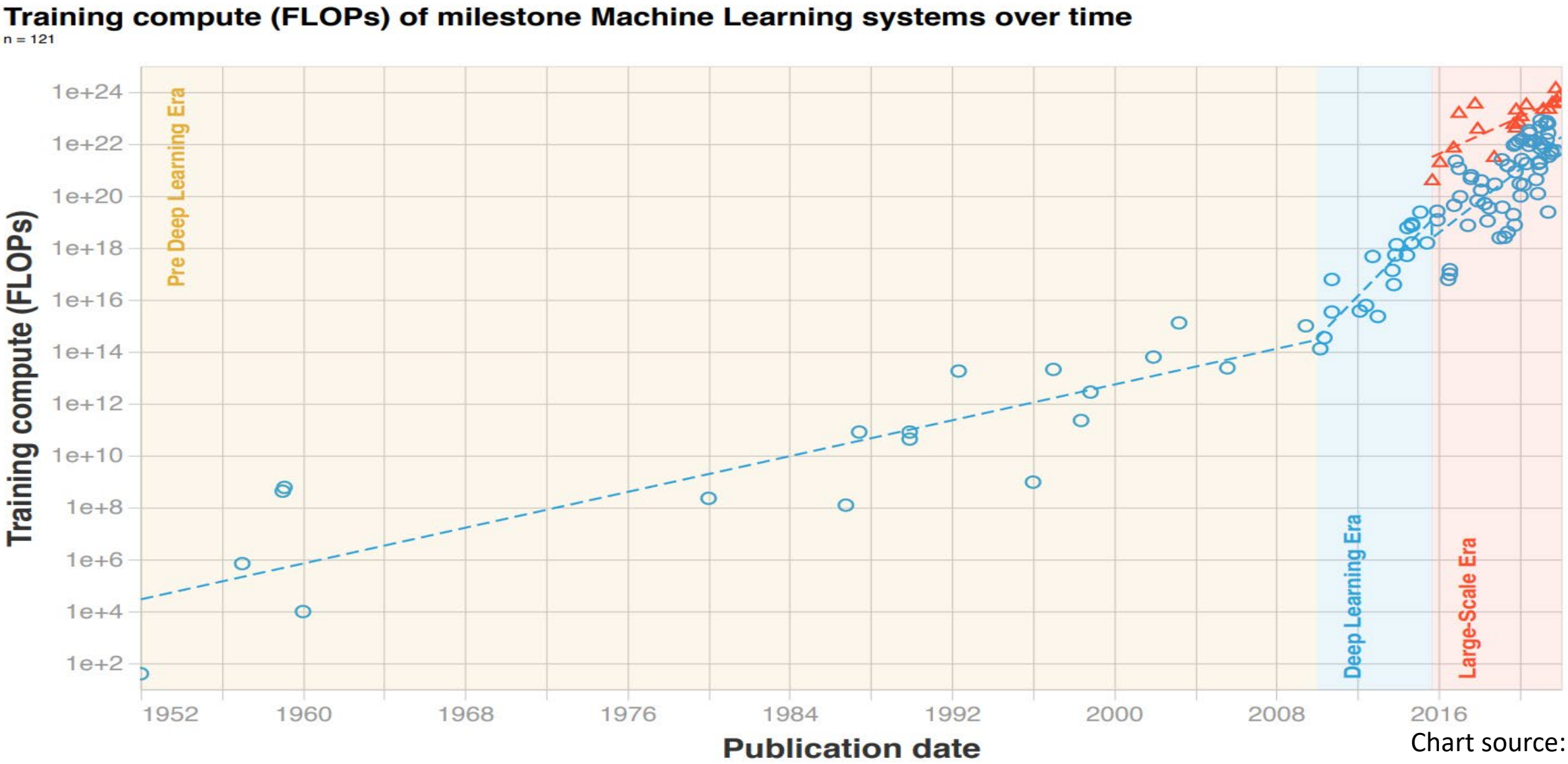


Chart source:
Seville et al 2023

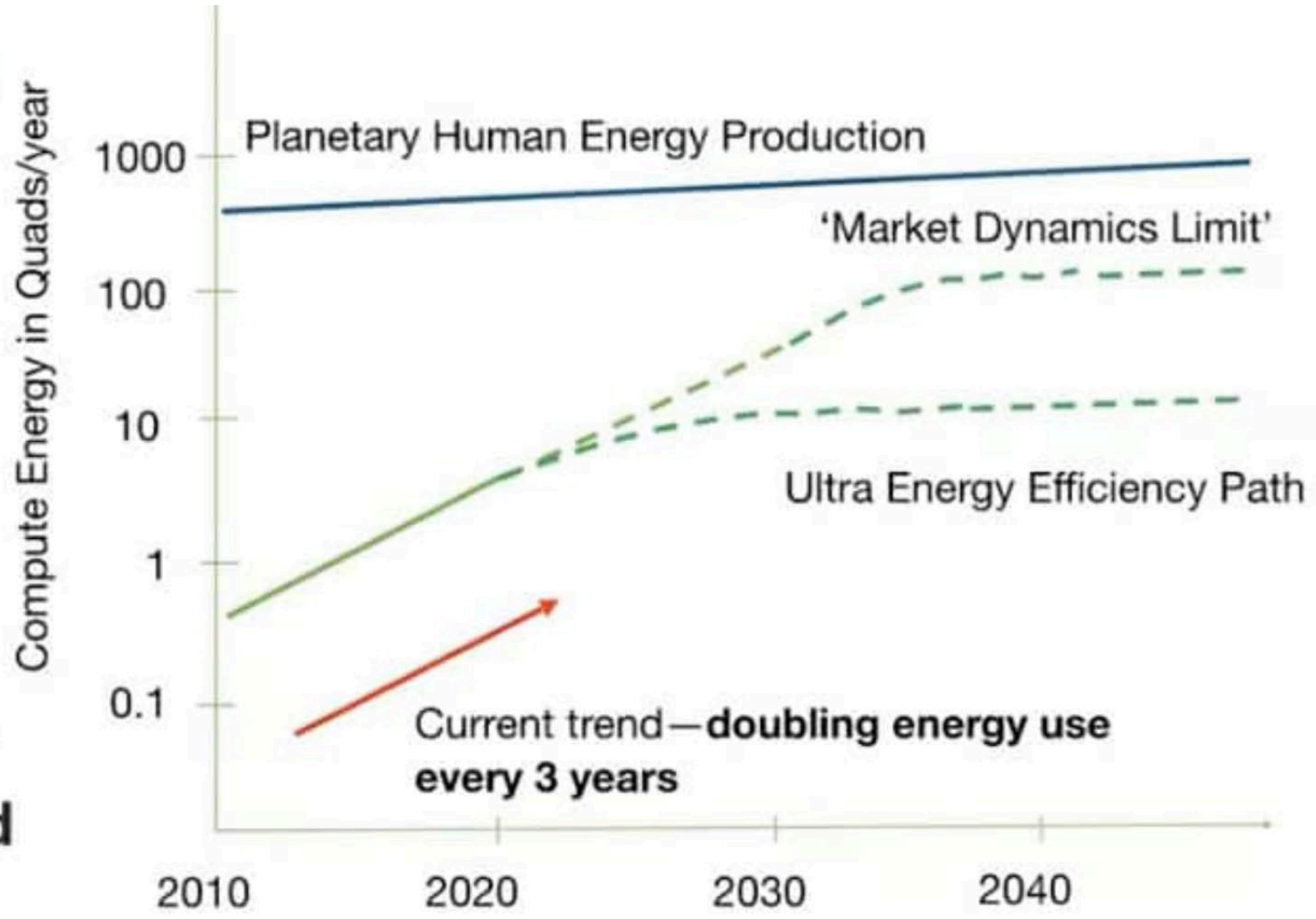
IEEE
IEEE photonics Society
semi
IEEE ELECTRONICS PACKAGING SOCIETY
ASME
ELECTRON DEVICES SOCIETY

Impact on Global Power Consumption of Microelectronics

Current trends would lead to market dynamics plateau at ~20% of planet energy—not so great for climate either

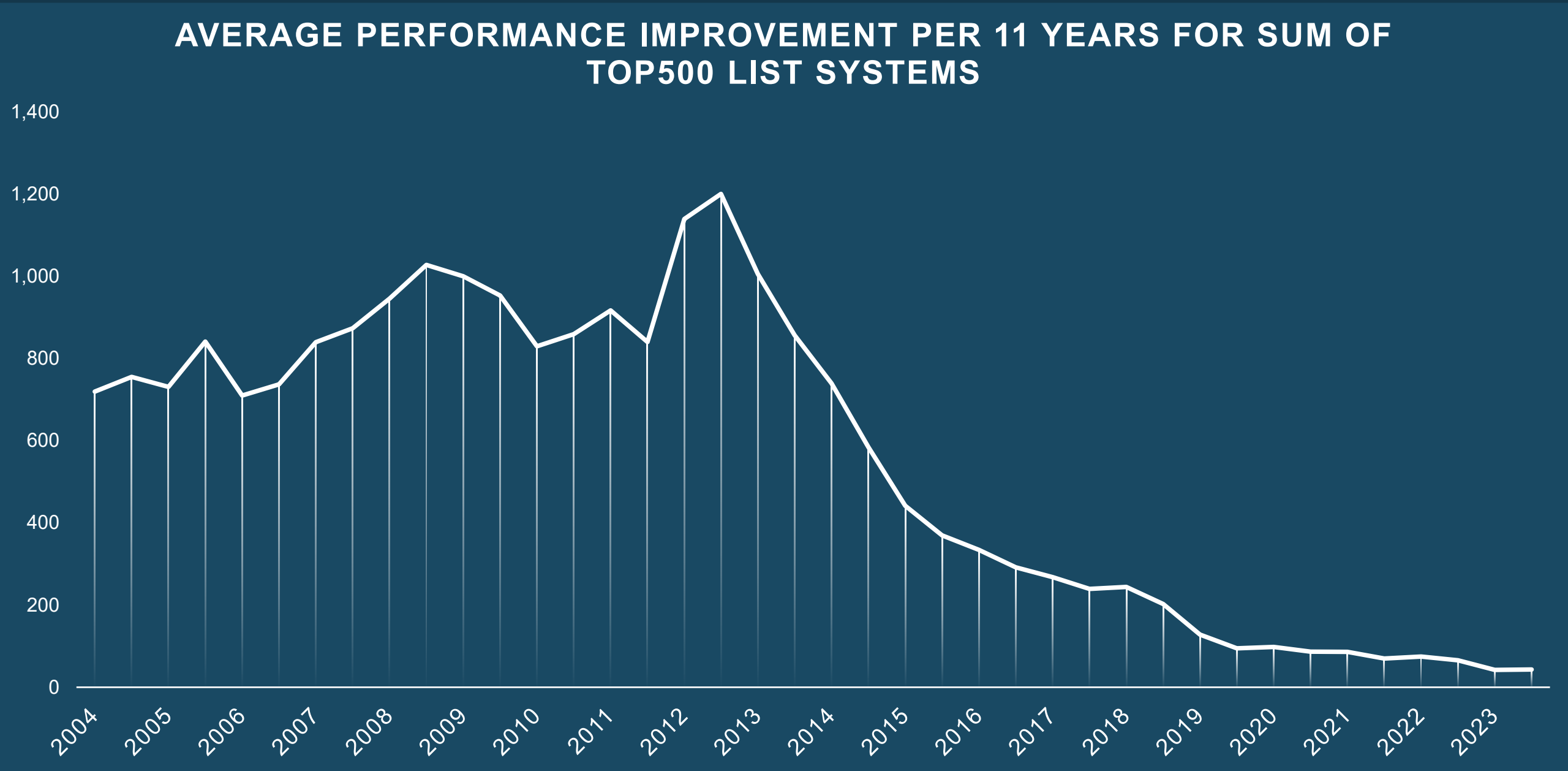
Market dynamics limit implies job losses and other negative economic impacts

Extreme energy efficiency is based on expanding and accelerating innovation



This is HPCs future if we continue business as usual!

… and "scale" might not be the answer…



Specialization:

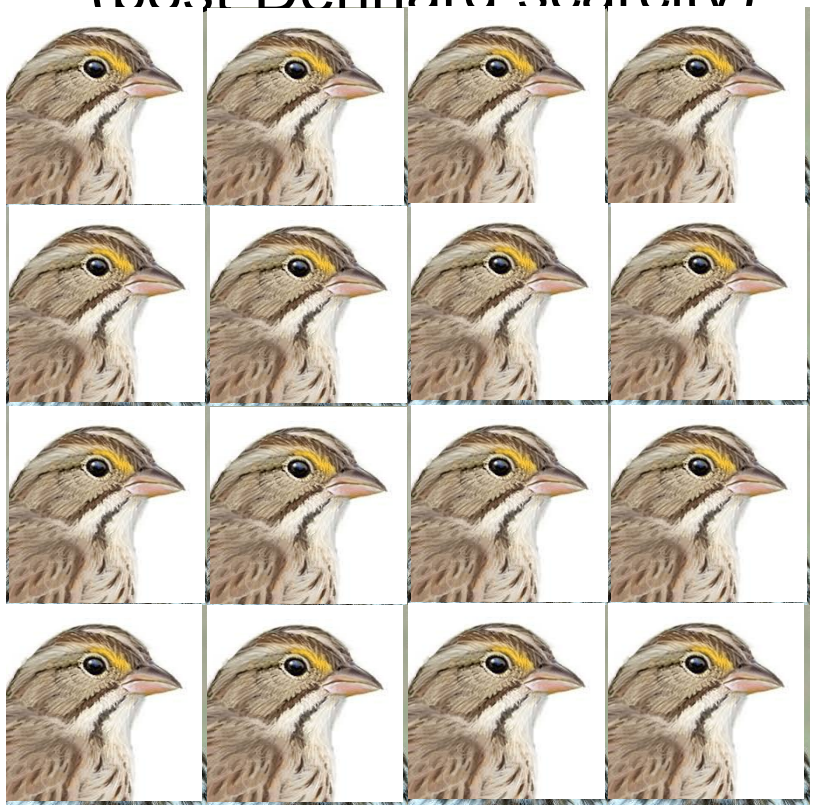
Natures way of Extracting More Performance in Resource Limited Environment

Powerful General Purpose

Xeon, Power

Many Lighter Weight

(post-Dennard scarcity)



KNL, AMD, Cavium/Marvell, GPU

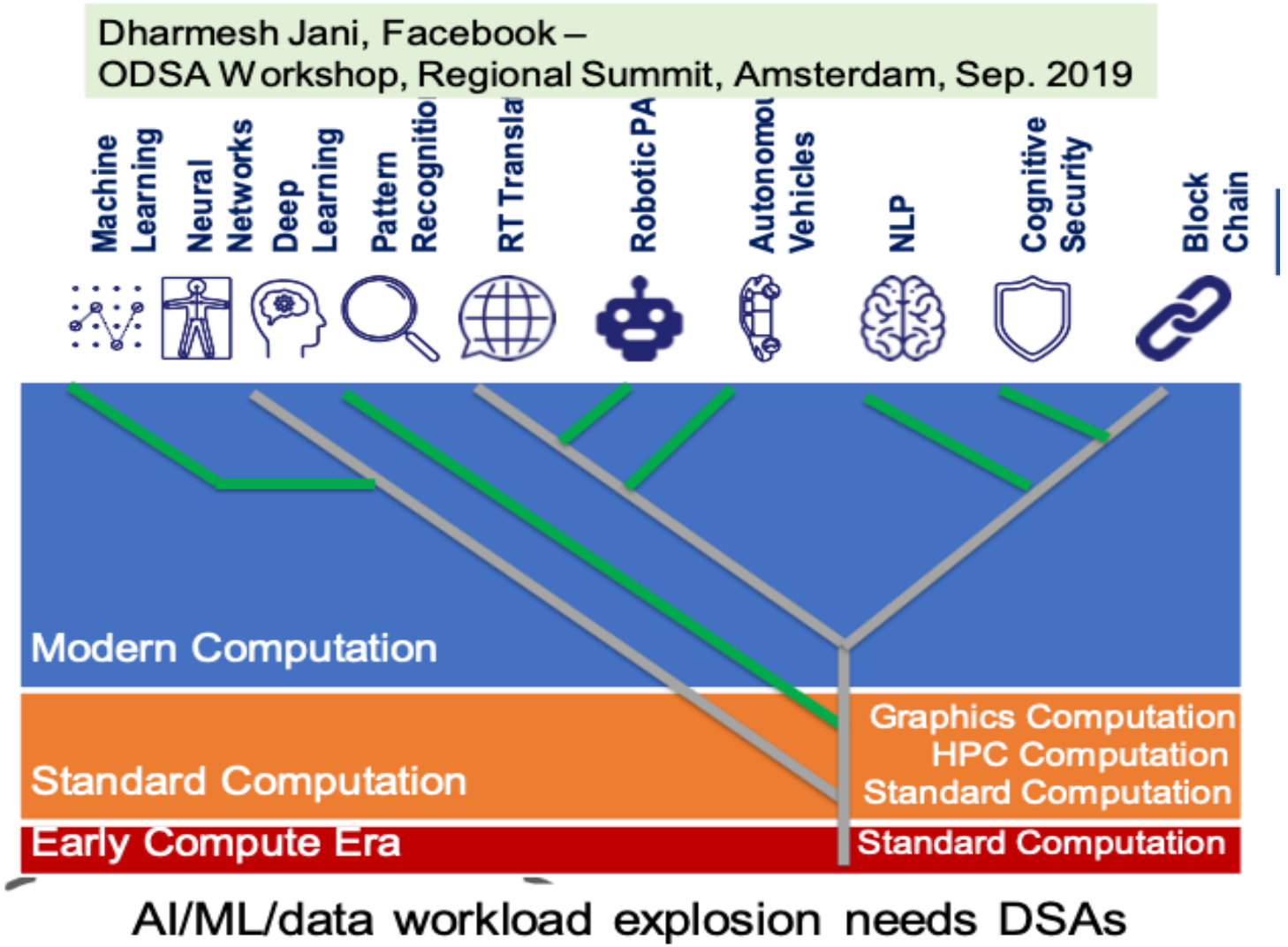
Many Different Specialized

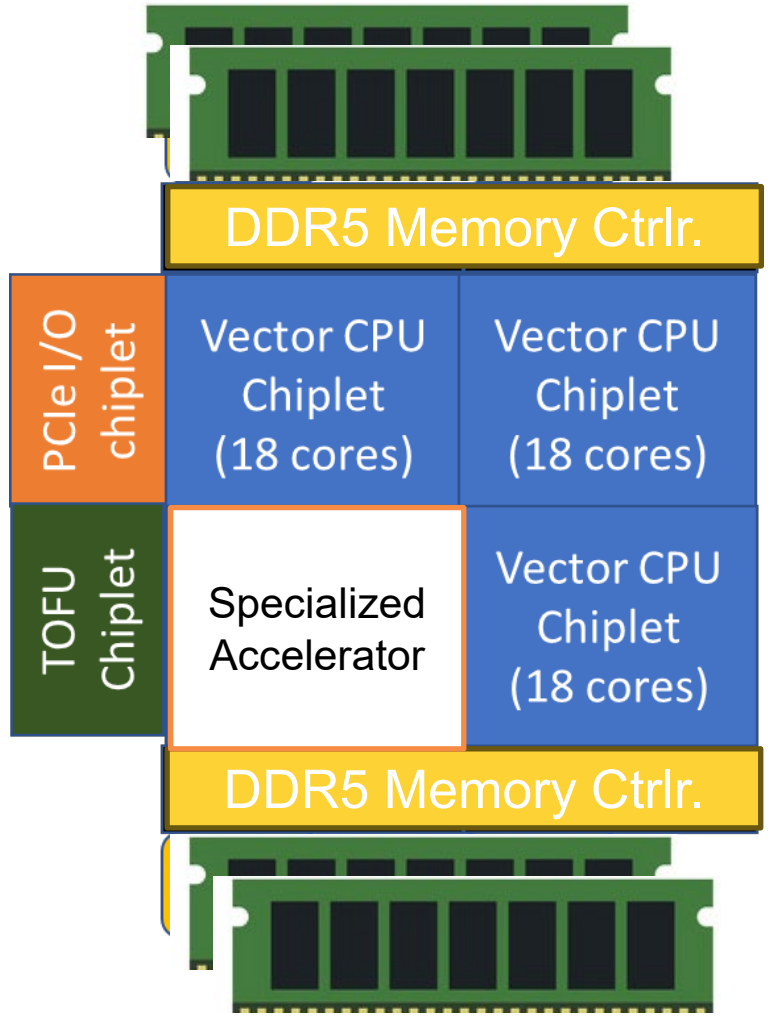
(Post-Moore Scarcity)

Apple, Google, Amazon Microsoft

How? Thought Experiment:

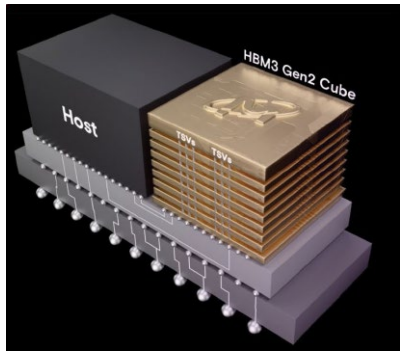
Chipletizing Fugaku With Modular Reusable Chiplets

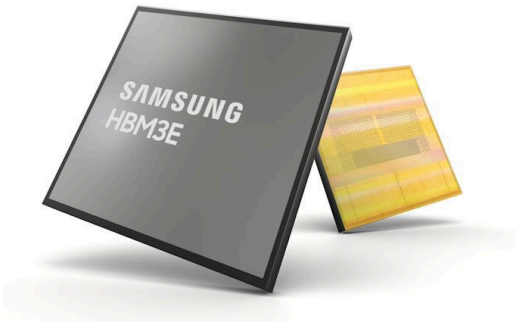




The Memory System, Interconnections and HI

- The memory system plays a critical role in terms of system-scale performance and energy consumption
- Increasingly larger datasets needs to be handled
- Inefficiencies in data transport have to be addressed
- Faster access to data is critical in AI/ML systems in particular
- **The 2024 edition of the HPC/Data Center will emphasize memory chiplets, advanced 3D interconnections, alternative memory system architectures and near-memory and in-memory processing paradigms**



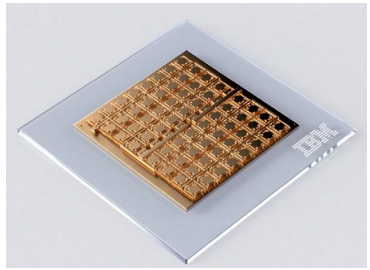


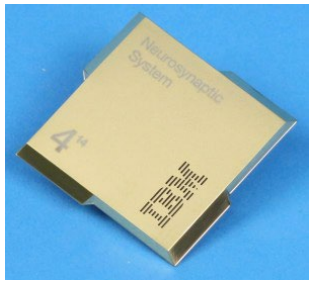
Pictures courtesy of Micron, Samsung and SK-Hynix

IEEE IEEE photonics Society semi IEEE ELECTRONICS PACKAGING SOCIETY ASME ELECTRON DEVICES SOCIETY

Alternative AI Compute Chiplets/Architectures

- **The 2024 edition of the HPC/Data Center will emphasize alternative chiplets and systems architecture for AI/ML, HPC/Data Centers**
  - Disaggregated/distributed processing architectures, interconnections
  - Analog chiplets for AI processing (ReRAMs, PCM crossbars etc.)
  - Chiplets for neuromorphic computing
  - Role of memory hierarchy
  - Recent innovations and trends in heterogeneously integrated HPC, AI, Data Center Products





Visuals courtesy of IBM, IBM, Intel and Brainchip-Edge Impulse; these are not HI products, but potential chiplets for HI

Other Updates Planned for the 2024 Edition

- Shorten the entire HPC/Data Center Chapter
- Replace Section 5 on QC with a shorter writeup and point to newly-formed QC chapter
- Emphasize the role of sustainable design and manufacturing practices, assembly and packaging and full life cycle-driven system development
- Update all tables
- Fork off chapter on Chiplets
  - Do a complete rewrite of the memory chiplet and interconnection sections
  - Chiplets chapter for HIR to focus on creating a viable open chiplets economy
- Update table at end of the chapter, update/add references
- Update other sections as needed

Questions?