JULY 9–13, 2023

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

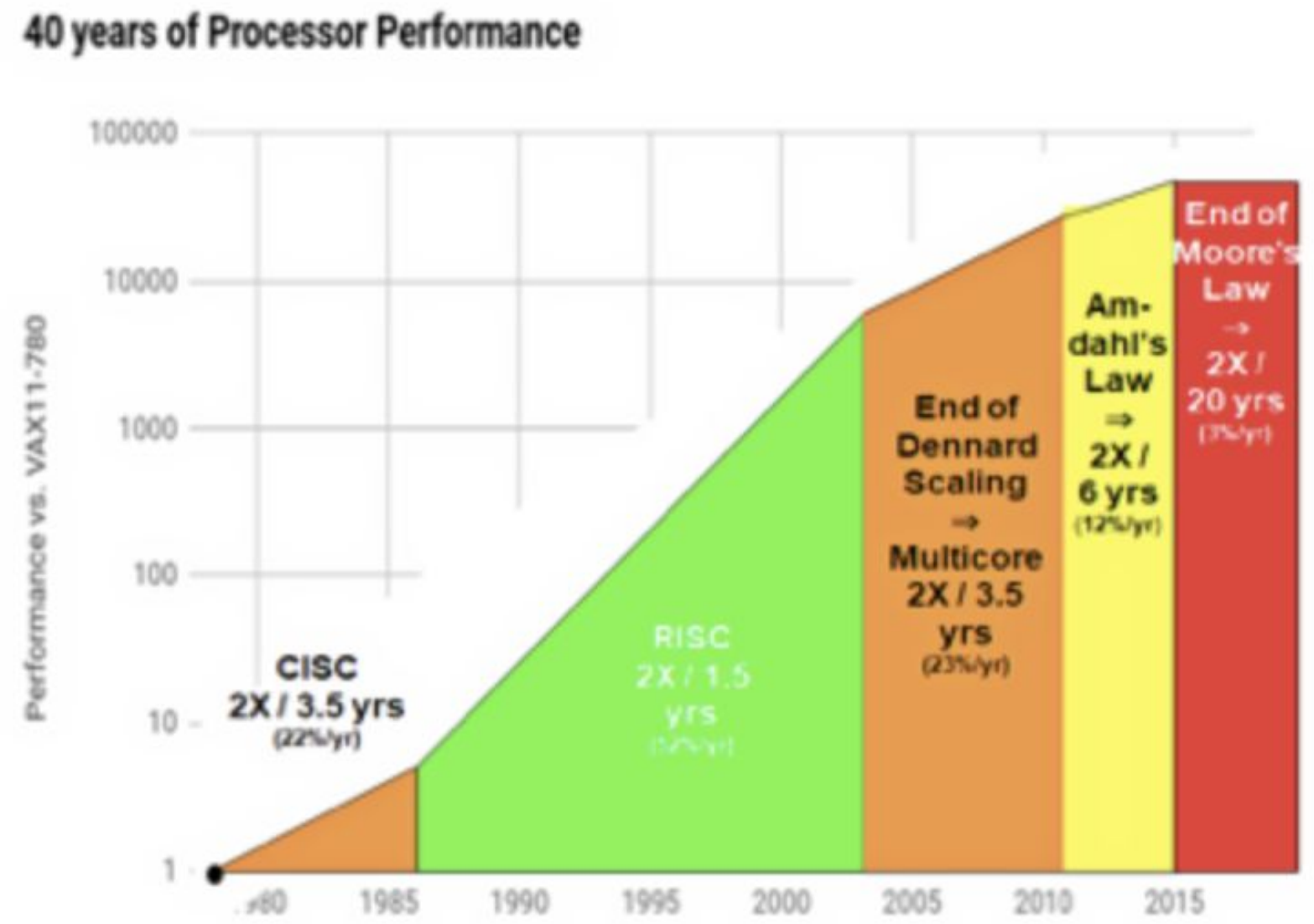
Slowing down of Moore's Law: How to Scale performance?

Durgesh Srivastava
NVIDIA Corp

Agenda

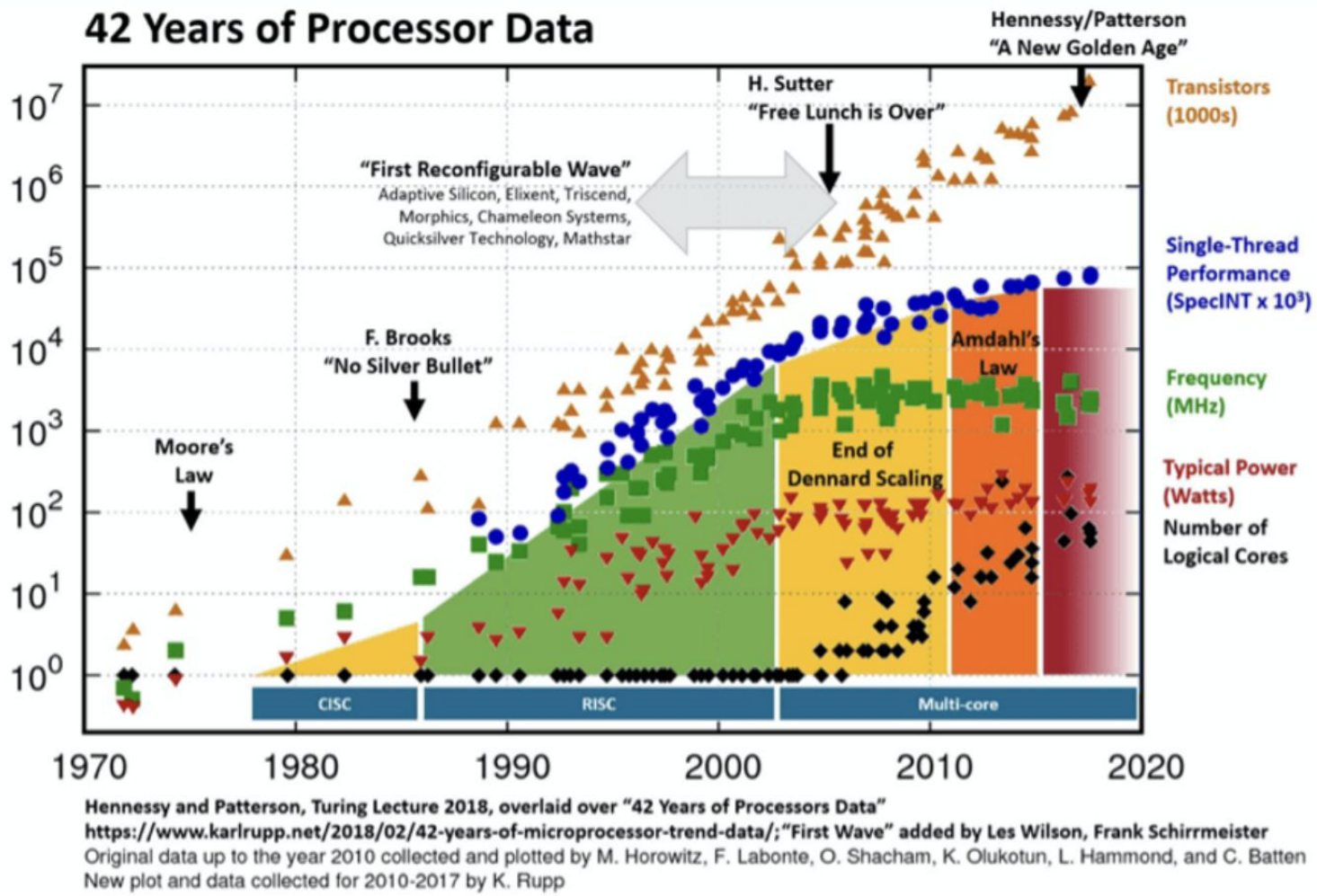
- Challenges
- Solutions
- Conclusions

Moore's Law: Difficult and Expensive



Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e. 2018

- Power Consumption
- Compute Scaling
- Security
- Yield/Cost
- Time to Market



Challenges for computing and technological advancement.

Unprecedented Demand Growth

The size of the **data and AI** market is on track to break the **$500 billion mark by 2024.**

Source: IDC Forecast Companies to spend 342 B on AI Solutions in 2021

Infrastructure and application modernization market potential is expected to reach more than **$500 billion by the end of 2025**

Source: Internal IDC & Google Cloud data, 2022

Partha Ranganathan (Google) @OCP 2022

Challenges for computing and technological advancement

Chiplets and Advance Packaging

- Full Circle: Disaggregation to Segregation to Disaggregation
  - Accelerators and Custom design
  - Heterogenous architectures: NVIDIA Grace + Hopper
  - Memory Pooling – avoiding data movements
- UCIe helps enabling a solution
  - EcoSystem Development
    - Chiplets
    - Fab and Testing
    - Integration and Functional Testing
    - Package
    - Production Testing
- Advancements
  - Hybrid and SOIC tech – Chiplets will look like monolithic by using advanced technology

Conclusions

Domain Specific Architectures and Accelerators

- CPU core helps with data pre-processing
- Custom design

Energy Efficiency and Thermal innovations

- Sustainable Energy usage
- Innovative Cooling solutions

Chiplet Ecosystem

- Packaging
- Easy integration

Memory Scaling and Security

- Minimize data movement
- Secure chiplet designs

BACKUP

UCIe: Key Metrics and Adoption Criteria

Key Performance Indicators

- Bandwidth density (linear & area)
  - Data Rate & Bump Pitch
- Energy Efficiency (pJ/b)
  - Scalable energy consumption
  - Low idle power (entry/exit time)
- Latency (end-to-end: Tx+Rx)
- Channel Reach
  - Technology, frequency, & BER
- Reliability & Availability
- Cost: Standard vs advanced packaging

Factors Affecting Wide Adoption

- Interoperability
  - Full-stack, plug-and-play with existing s/w
  - Different usages/segments - ubiquity
- Technology
  - Across process nodes & packaging options
  - Power delivery & cooling
  - Repair strategy (failure/yield improvement)
  - Debug – controllability & observability
- Broad industry support / Open ecosystem
  - Learnings from other standards efforts

UCIe is architected and specified from the ground-up to deliver the best KPIs while meeting wide adoption criteria

UCIe 1.0: Characteristics and Key Metrics

CHARACTERISTICS	STANDARD PACKAGE	ADVANCED PACKAGE	COMMENTS
Data Rate (GT/s)	4, 8, 12, 16, 24, 32		Lower speeds must be supported -interop (e.g., 4, 8, 12 for 12G device)
Width (each cluster)	16	64	Width degradation in Standard, spare lanes in Advanced
Bump Pitch (um)	100 – 130	25 - 55	Interoperate across bump pitches in each package type across nodes
Channel Reach (mm)	<= 25	<=2	

KPIs / TARGET FOR KEY METRICS	STANDARD PACKAGE	ADVANCED PACKAGE	COMMENTS
B/W Shoreline (GB/s/mm)	28 – 224	165 – 1317	Conservatively estimated: AP: 45u; Standard: 110u; Proportionate to data rate (4G – 32G)
B/W Density (GB/s/mm^2)	22-125	188-1350	
Power Efficiency target (pJ/b)	0.5	0.25	
Low-power entry/exit latency	0.5ns <=16G, 0.5-1ns >=24G		Power savings estimated at >= 85%
Latency (Tx + Rx)	< 2ns		Includes D2D Adapter and PHY (FDI to bump and back)
Reliability (FIT)	0 < FIT (Failure In Time) << 1		FIT: #failures in a billion hours (expecting ~1E-10) w/ UCIe Flit Mode

UCIe 1.0 delivers the best KPIs while meeting the projected needs for the next 5-6 years.
Wide industry leader adoption spanning semiconductor, manufacturing, assembly, & cloud segments.