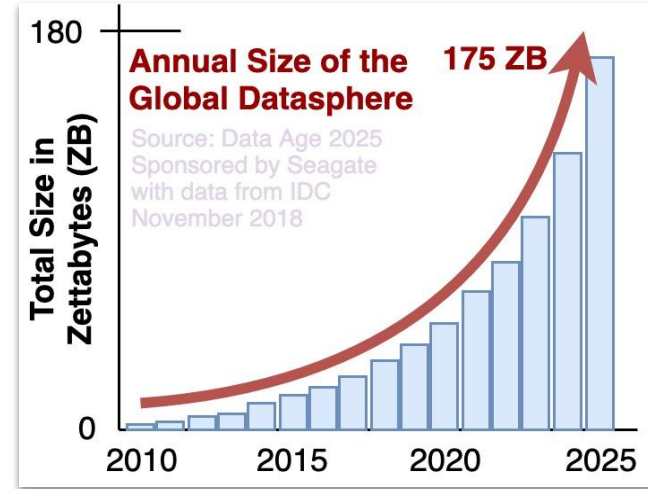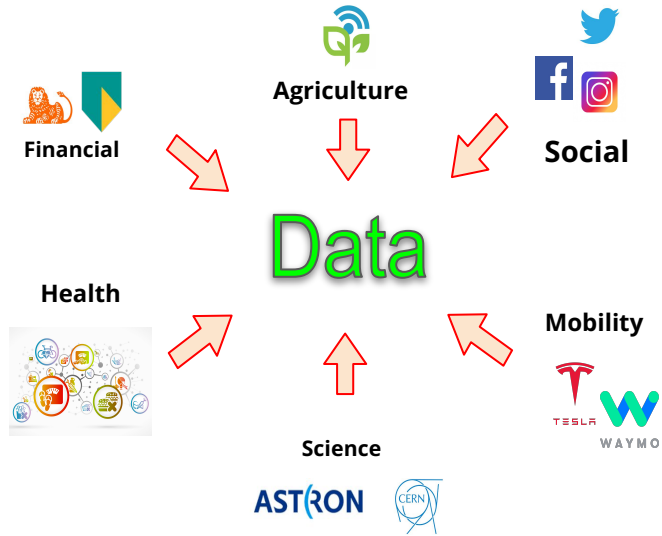# Hyperion: A Unified, Zero-CPU Data-Processing Unit (DPU)
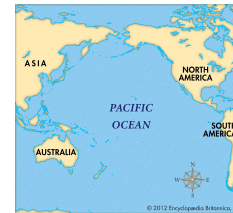
Marco Spaziani Brunella, Marco Bonola and **Animesh Trivedi**

CompSys 2022

The Data Explosion
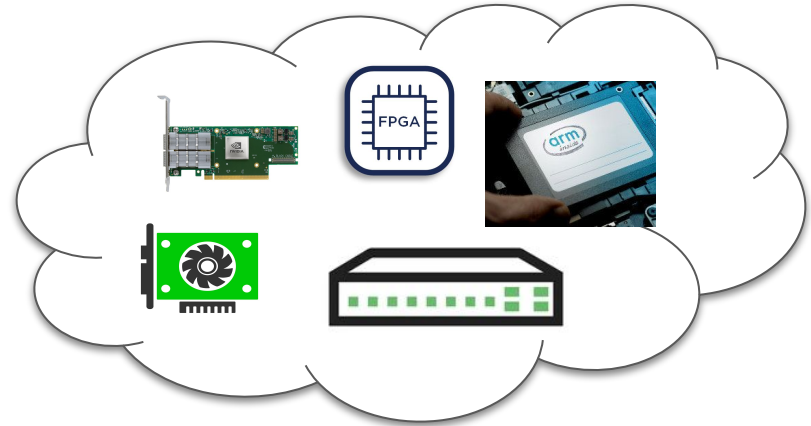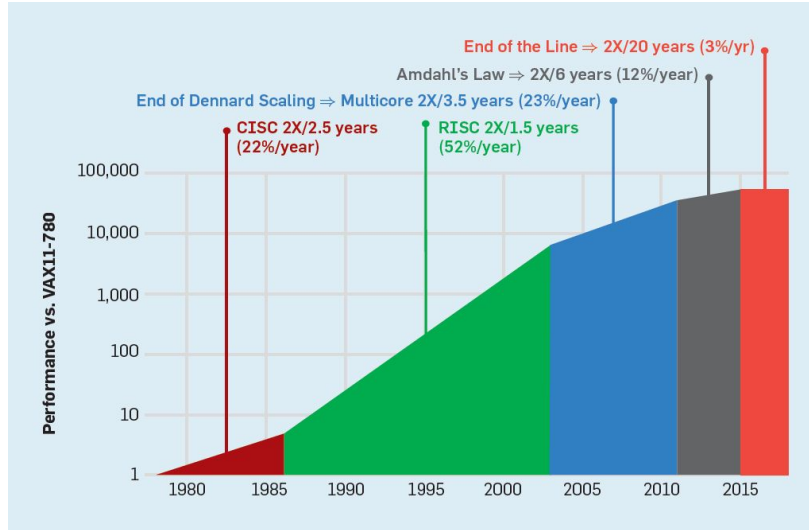
Financial

Agriculture

Social

Health

Data

Mobility

Science

Annual Size of the Global Datasphere

175 ZB

Source: Data Age 2025
Sponsored by Seagate
with data from IDC
November 2018

Total Size in Zettabytes (ZB)

200 Zettabytes *revised*

Fills the Pacific Ocean 200x over

CPU - as the Performance Horse

End of the Line ⇒ 2X/20 years (3%/yr)
Amdahl's Law ⇒ 2X/6 years (12%/year)
End of Dennard Scaling ⇒ Multicore 2X/3.5 years (23%/year)
CISC 2X/2.5 years (22%/year)
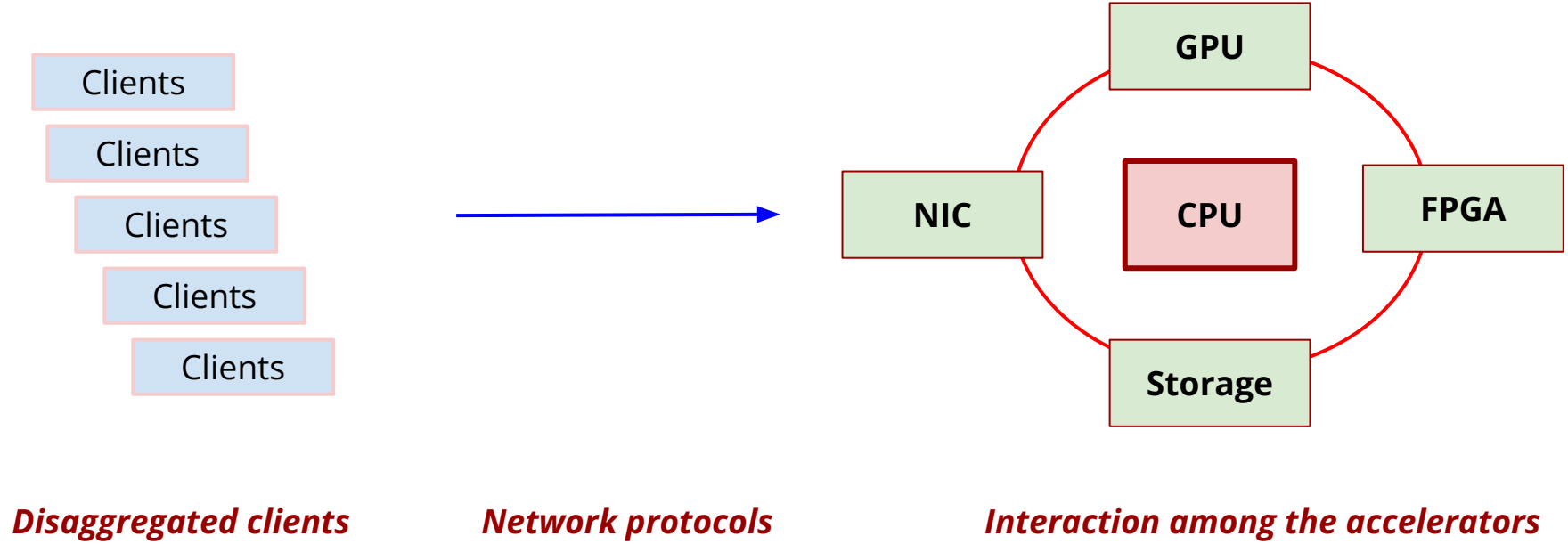RISC 2X/1.5 years (52%/year)

- **Stalling** of Moore's Law and Dennard Scaling
- Turing Tax - **the cost of Generalization**
- **Security** considerations
- **Energy needs**
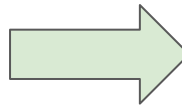
Rise of accelerator-centric computing

3

Imagine this setup

Clients

Clients

Clients

Clients

Clients

GPU

NIC

CPU

FPGA

Storage

Disaggregated clients

Network protocols

Interaction among the accelerators

4

The Key Challenge with the CPU in the Loop

1. The CPU controls the control path and resource allocation
   a. Coordinate control flow among accelerators - which buffers to allocate, pin, DMA
   b. Control the data transfer among accelerators - when to initiate and how to initiate
   c. Done with pair-wise accelerator integrations, but multiple?

2. The CPU dictates the computing abstractions
   a. Shared memory, virtual memory, processes, context switches, files
   b. ***Keeping the memory coherent between the host's view and accelerator view***

3. The CPU dictates the innovation and imagination
   a. Active and passive disaggregation
   b. Designing a new interconnect, network discovery protocols
   c. Scalable energy needs

Hyperion: A Zero-CPU Data Processing Unit (DPU)

Hardware:

- FPGA + NIC + Storage = DPU

Software:

- A new compiler
- eBPF as an **IR** for *(any)* hardware

Client:

- *Disaggregated clients*
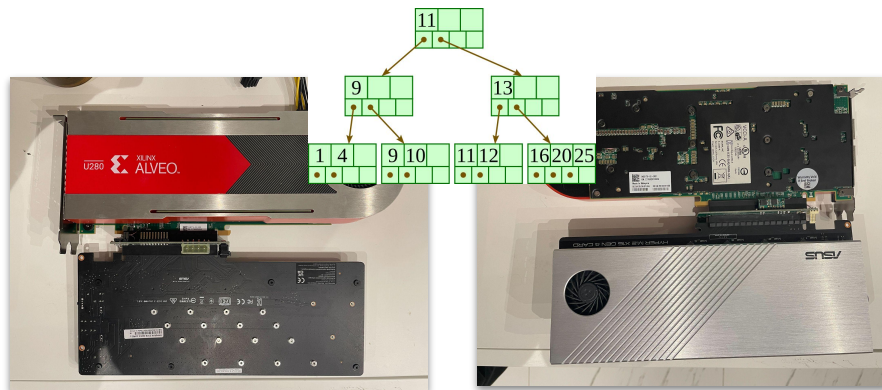- Network protocols - NVMoF
- Application-level, KV, NFS, DSes

LLVM

new_compiler
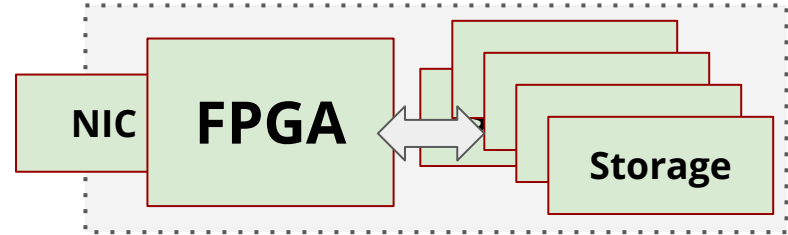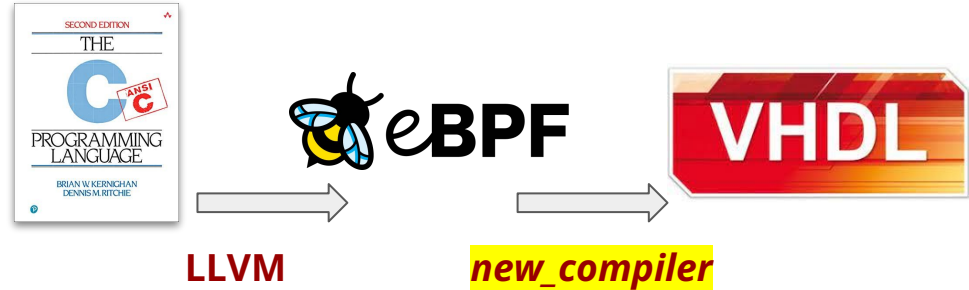
Where are we going from here?

Hyperion: A Case for Unified, Self-Hosting, Zero-CPU Data-Processing Units (DPUs):

- https://arxiv.org/abs/2205.08882

Backup slides

Related Work

What	Examples
Net + Accel	SmartNICs [5, 110], AcclNet [53], hXDP [35]
Net + GPU	GPUDirect [102], GPUNet [78]
Sto + GPU	Donard [22], SPIN [25], GPUfs [124], GPUDirect [103], nvidia BAM [113]
Net + Sto	iSCSI, NVMoF (offload [117], BlueField [5]), i10 [68], ReFlex [80]
Sto + Accel	ASIC/CPU [60, 83, 121], GPUs [25, 26, 124], FPGA [69, 116, 119, 143], Hayagui [15]
Hybrid System	with ARM SoC [3, 47, 90], BEE3 [44], hybrid CPU-FPGA systems [39, 41]
DPUs	Hyperion (stand-alone), Fungible (MIPS64 R6 cores) DPU processor [54], Pensando (host-attached P4 Programmable processor) [108], BlueField (host-attached, with ARM cores) [5]