

# Computing - Big Impact in the 21<sup>st</sup> Century



Wen-mei Hwu

Professor and Sanders-AMD Chair, ECE  
University of Illinois at Urbana-Champaign



1988

Start of the Hwu Family

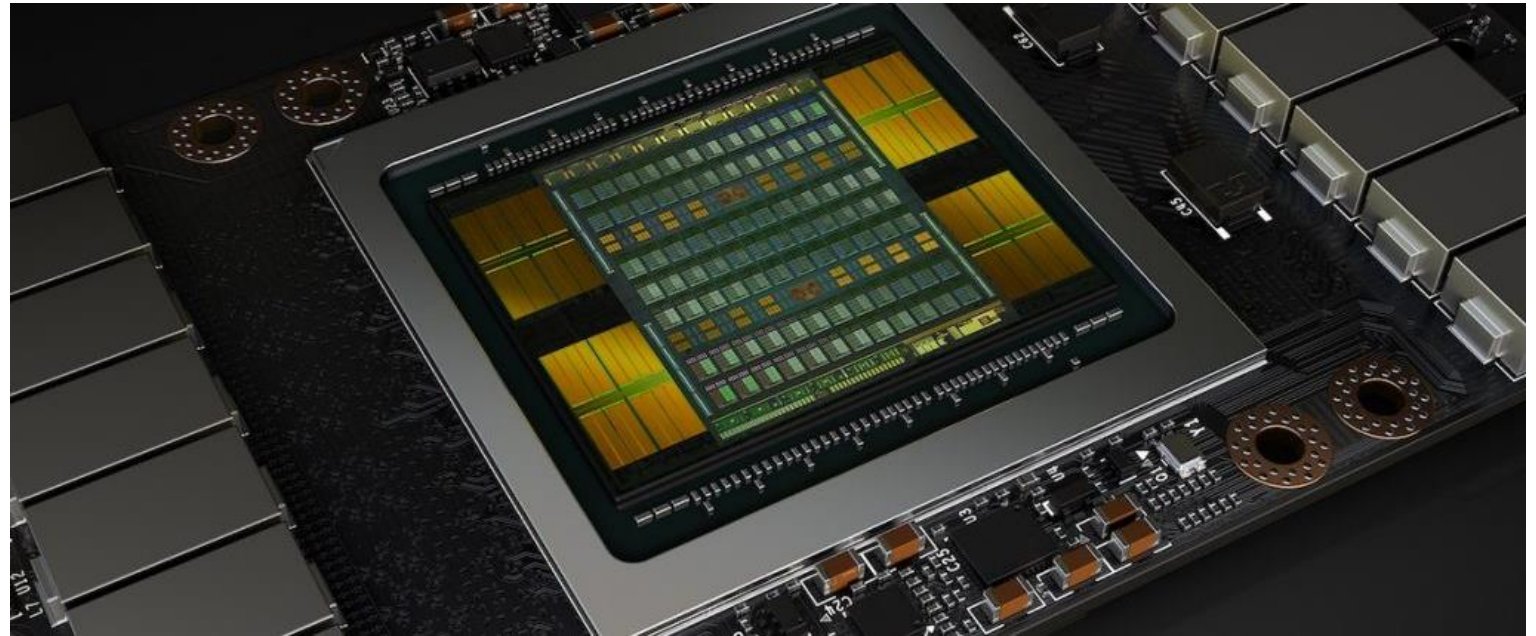


2016

Yale Wins Franklin Medal

Int 286

A80286-8  
L9409047  
© 1982, 1985



134K vs. 12.1B transistors

12 MHz vs. 1.1 GHz

1.5  $\mu\text{m}$  vs. 12 nm

2.7 MIPS (needs 287 for FP) vs. 14 TFLOPS

---

1MB DRAM vs. 16GB HBM

# The Industry Landscape

- Apple II
  - Sony, DEC, IBM, Intel and Microsoft
  - ...
- iPhone X
  - Samsung, Apple, NVIDIA, Amazon, Google, and Facebook
  - ...


# Innovation

A high-value concept in the right  
historical context

# Important Innovations in Recent History

- Telescope
- Microscope
- Electricity
- Telephone
- Medical imaging
- Electrical Motor
- Automobiles
- Airplane
- Credit cards
- Radar
- Clean Energy
- Mobile phones
- Internet and search engine
- eCommerce
- Social media
- GPS

Future innovations will rely  
heavily on **computing**

A large crowd of people is seated in the background, looking intently at two large monitors in the foreground. The monitors display a chess match. The left monitor shows a chessboard with the moves '1. e4 c6 2. d4 d5 3. Nc3 dxe4' at the bottom. The right monitor shows a close-up of a person's hands moving a chess piece on a board.

On May 11, 1997, IBM Deep Blue  
defeated world champion of chess  
(Garry Kasparov)

1. e4 c6 2. d4 d5 3. Nc3 dxe4



Feb 16, 2011, IBM Watson  
defeated two world  
champions of Final Jeopardy!


Score	Question	Answer
\$300,000	Who is Stoker?	Stoker
\$1,000	(FOR ONE WELL-KNOWN NEW COMPUTER OVERLOADS)	Stoker
\$11,000,000	Who is Stoker?	Stoker
\$17,973	Who is Stoker?	Stoker
\$200,000	Who is Stoker?	Stoker
\$5600	Who is Stoker?	Stoker



On March 15, 2016, Google AlphaGo  
defeated South Korean  
Go grandmaster Lee Sedol



LEE SEDOL

A photograph of a man in a blue long-sleeved shirt sitting at a poker table, resting his chin on his hand in a thoughtful expression. In the background, another man in a dark suit and patterned tie is visible, looking towards the camera. The setting appears to be a poker tournament, with a "PITTSBURGH POKER OPEN" sign and a green and white poster visible in the background.

In Jan 2017, CMU Libratus  
beat professional players in heads-up  
no-limit Texas hold'em poker game



On Jan 24, 2019, Google AlphaStar  
defeated human pros at real-time  
strategy game StarCraft II

intelligence<sup>2</sup>  
DEBATES

#iq2uslive

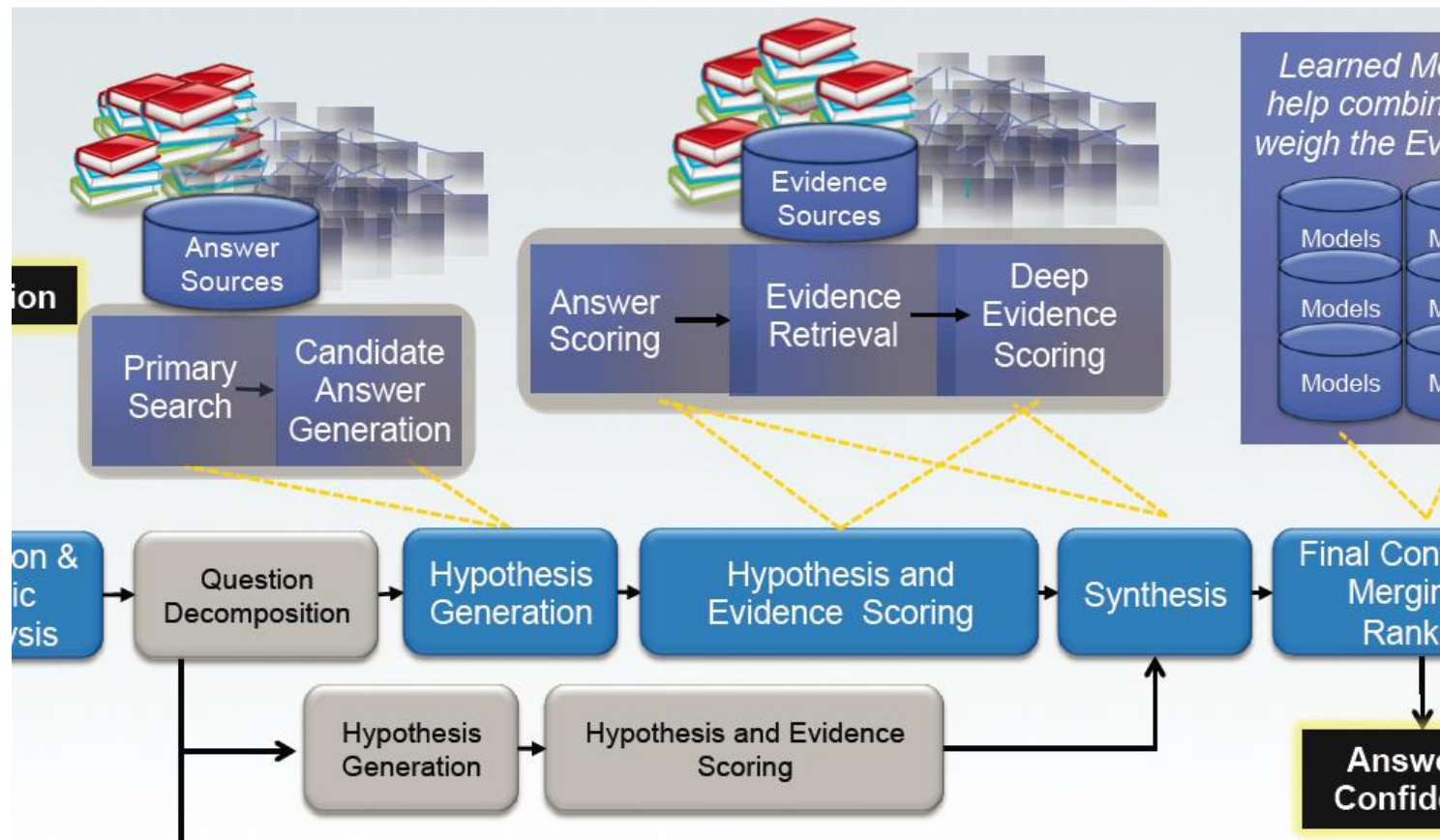
On Feb 11, 2019, IBM Project Debater  
debated with an European world  
champion





## Hardware for Watson Jeopardy! 2011

- 90 x IBM Power 750 servers
  - 2880 Power7 cores
  - 3.55 GHz clock
  - 80 TeraFLOPS
- 15 Terabytes of DRAM
- 20 Terabytes of disk
- 10 Gb Ethernet network
- > 100,000 Watt power



“Watson DeepQA generates and scores many hypotheses using an **extensible collection** of Natural Language Processing, Machine Learning and Reasoning Algorithms. These gather and weigh evidence over both unstructured and structured content to determine the answer with the best confidence.”

# Software for Watson Jeopardy! 2011

# Novelty vs. Great Product

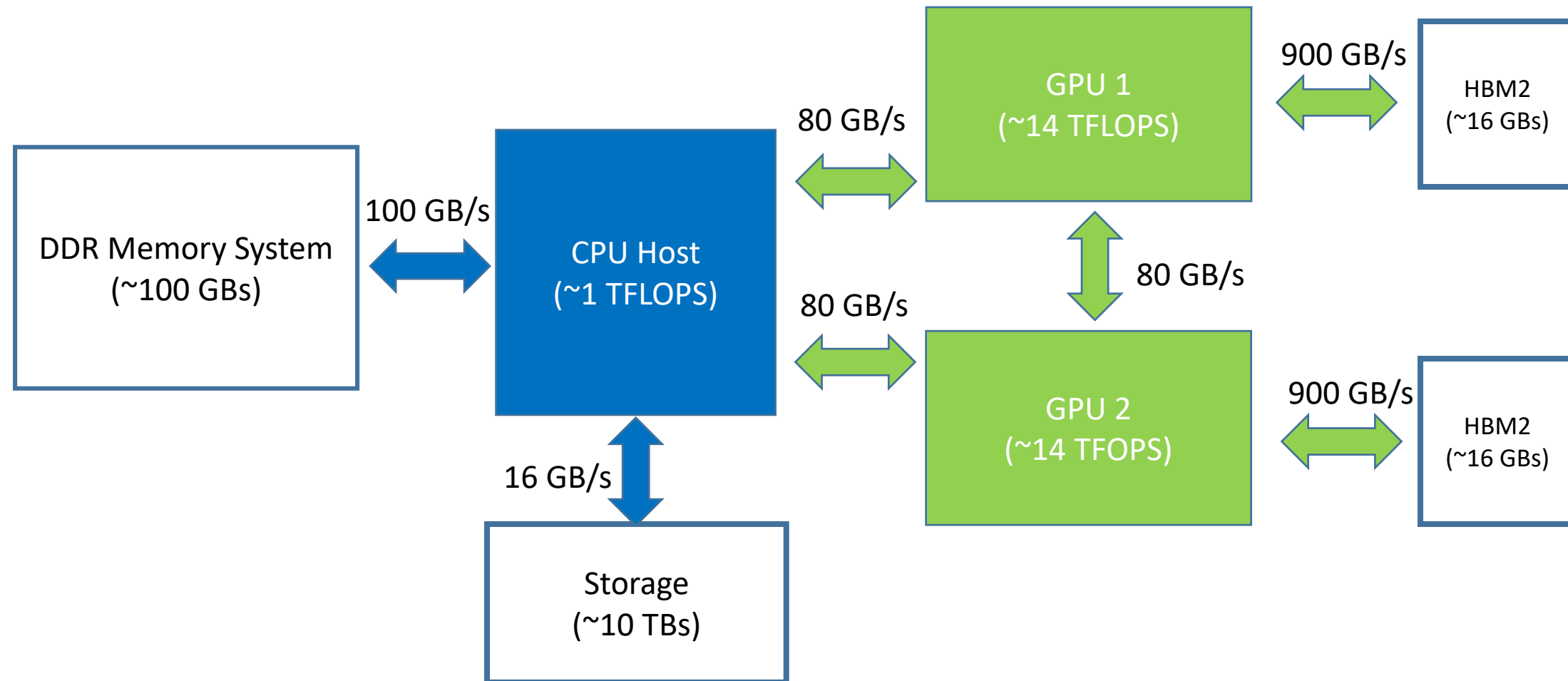


German *Flocken Elektrowagen* of 1888, regarded as the first electric car of the world

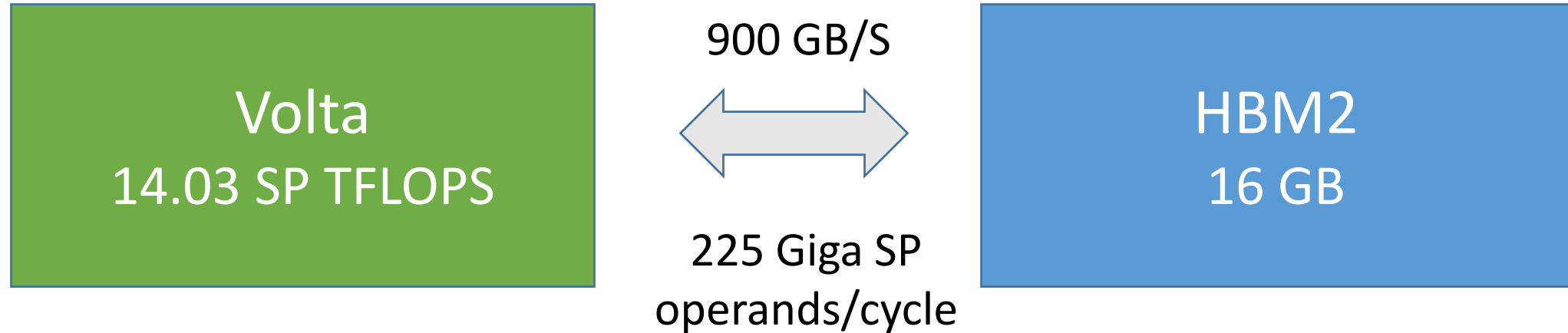


American Tesla Model X of 2017, whose producer is worth more than GM and Ford

# A Simplified View of IBM Newell with NVIDIA Volta GPUs



# Data Access Challenge (HBM2)

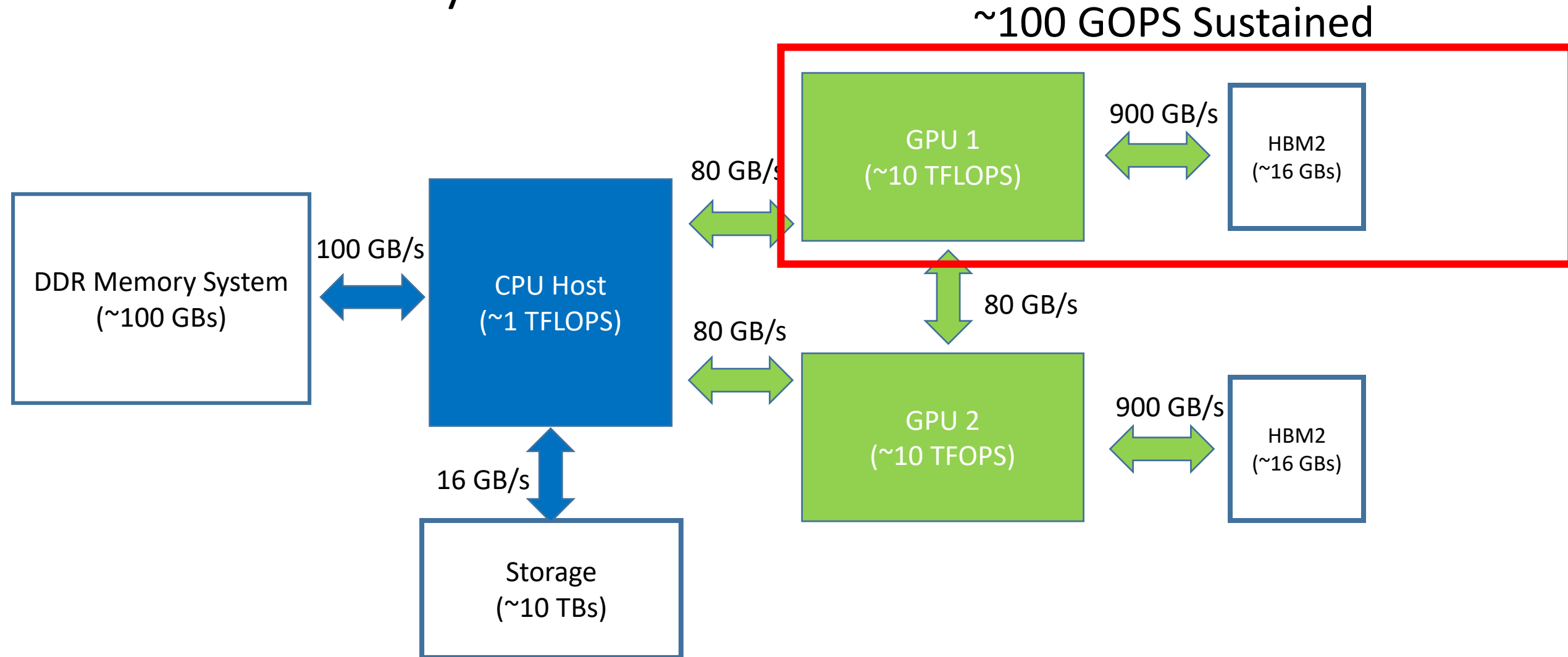


Each operands must be used **62.3 times** once fetched  
to achieve peak FLOPS rate.

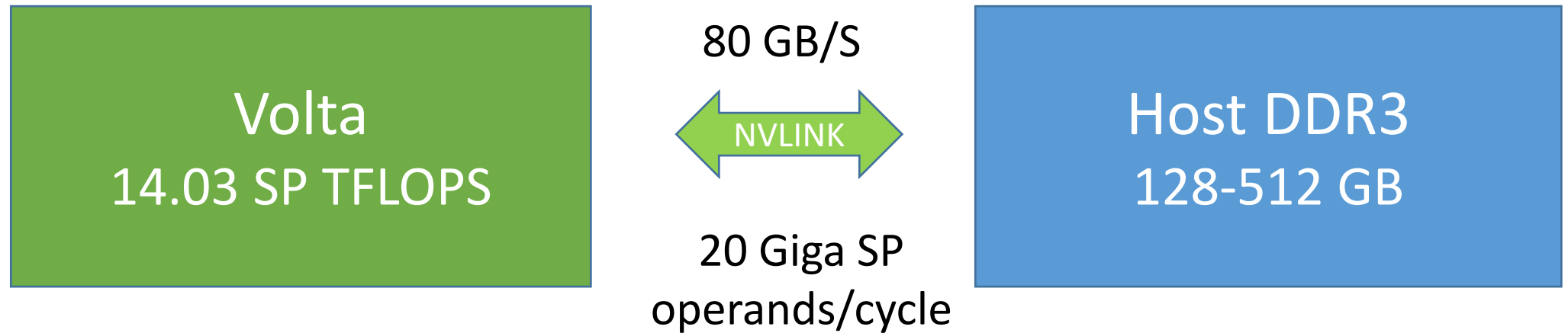
or

Sustain < **1.6%** of peak without data reuse

# Graph Analytics Example – if graph fits into GPU Memory



# Data Access Challenge (Host DDR3)

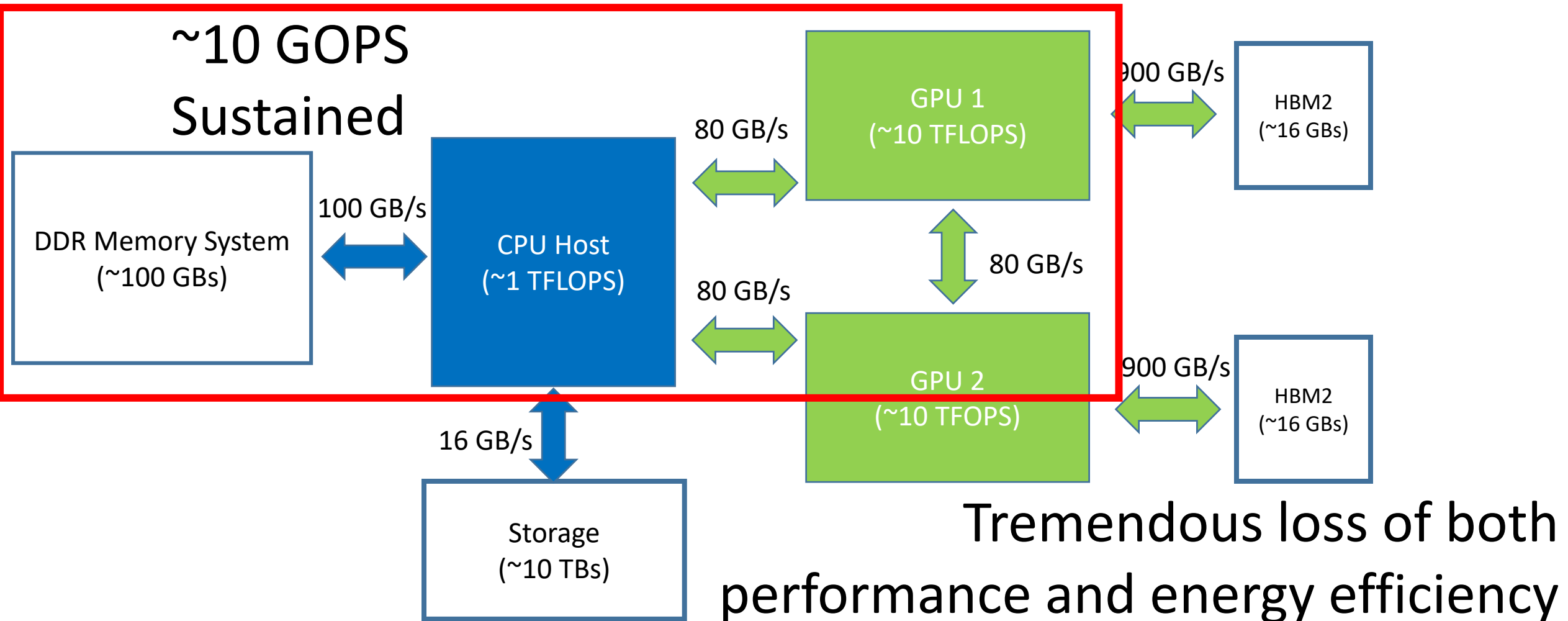


Each operands must be used **700 times** once fetched  
to achieve peak FLOPS rate.

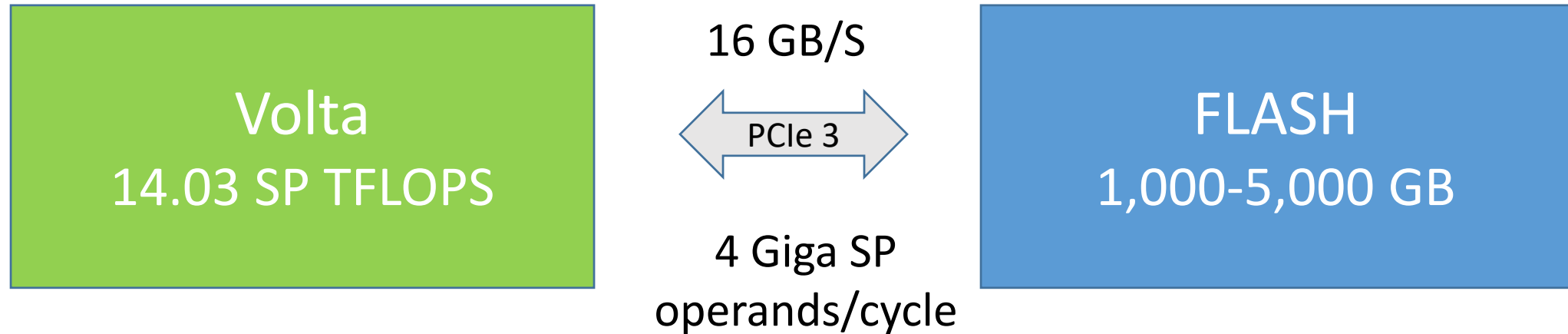
or

Sustain < **0.14% peak** without data reuse

# Graph Analytics Example – if graph fits into Host DDR Memory



# Data Access Challenge (FLASH SSD)

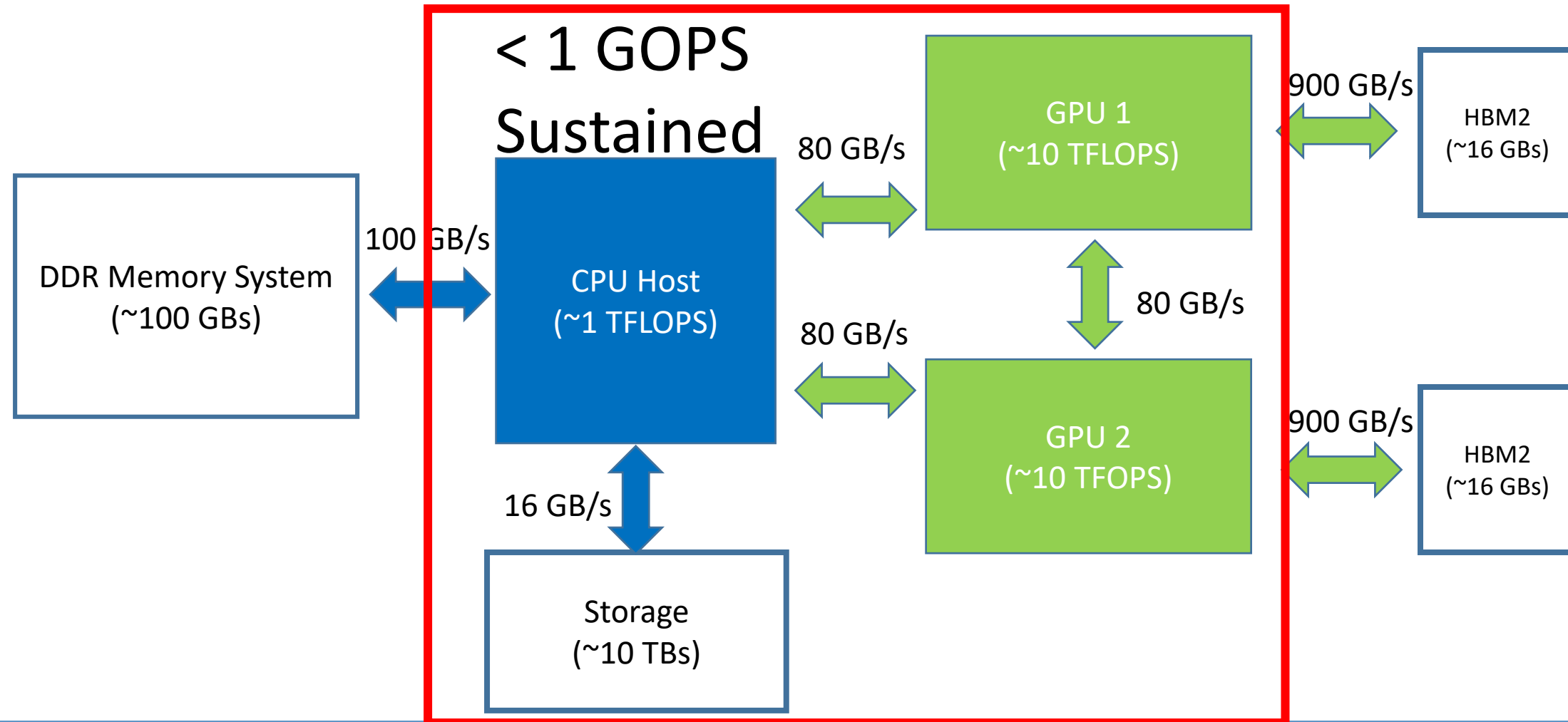


Each operands must be used **3,507 times** once fetched  
to achieve peak FLOPS rate.

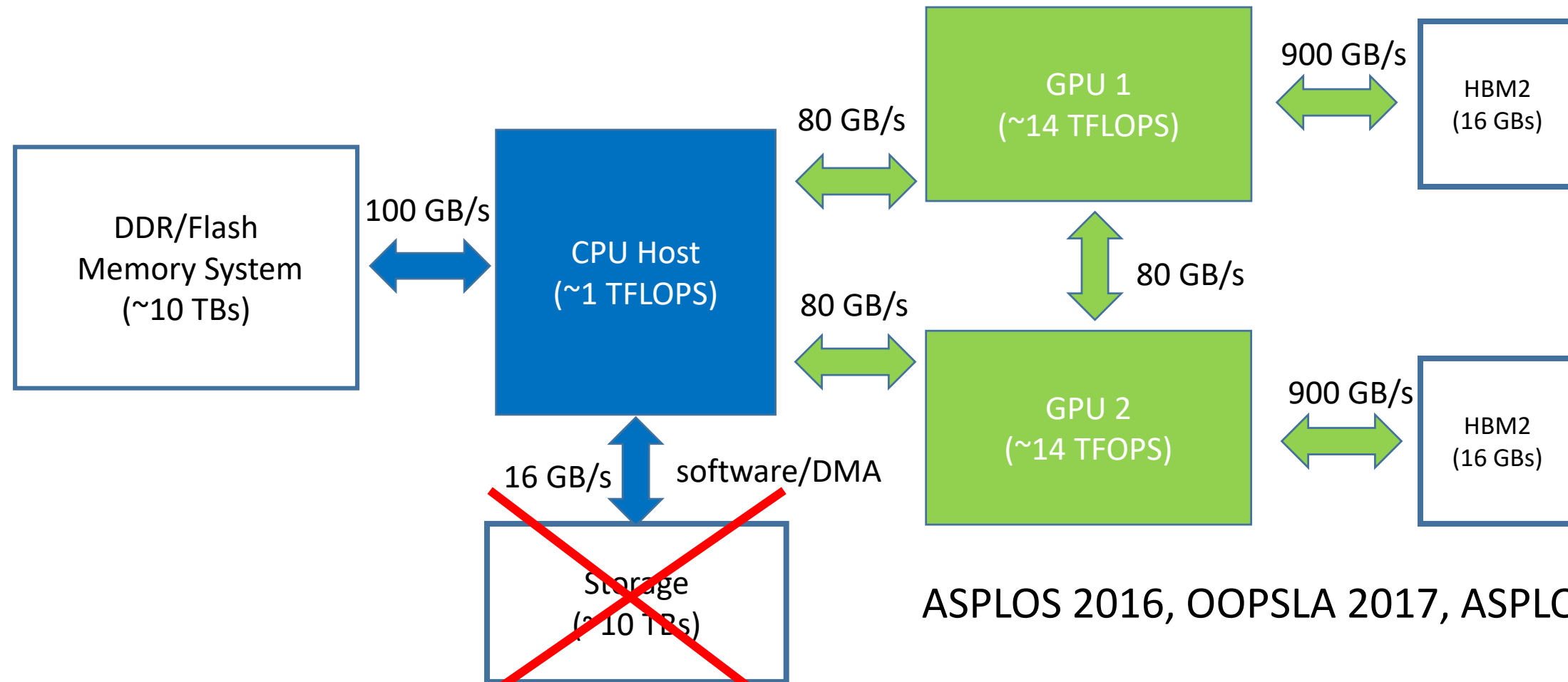
or

Sustain **< 0.03%** of peak without data reuse

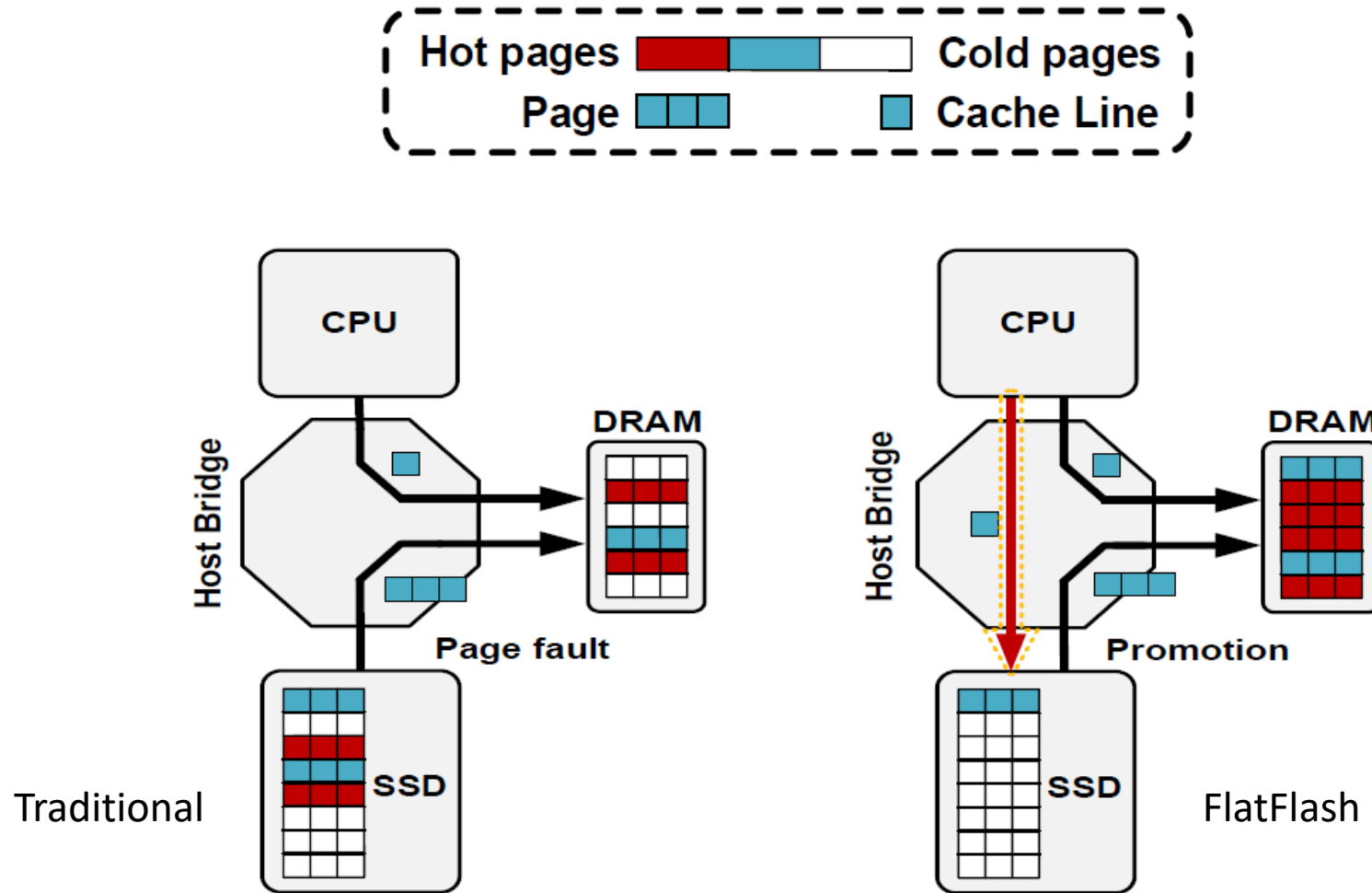
# Graph Analytics Example – if graph is accessed from storage



# Erudite Vision: remove file system from data access path

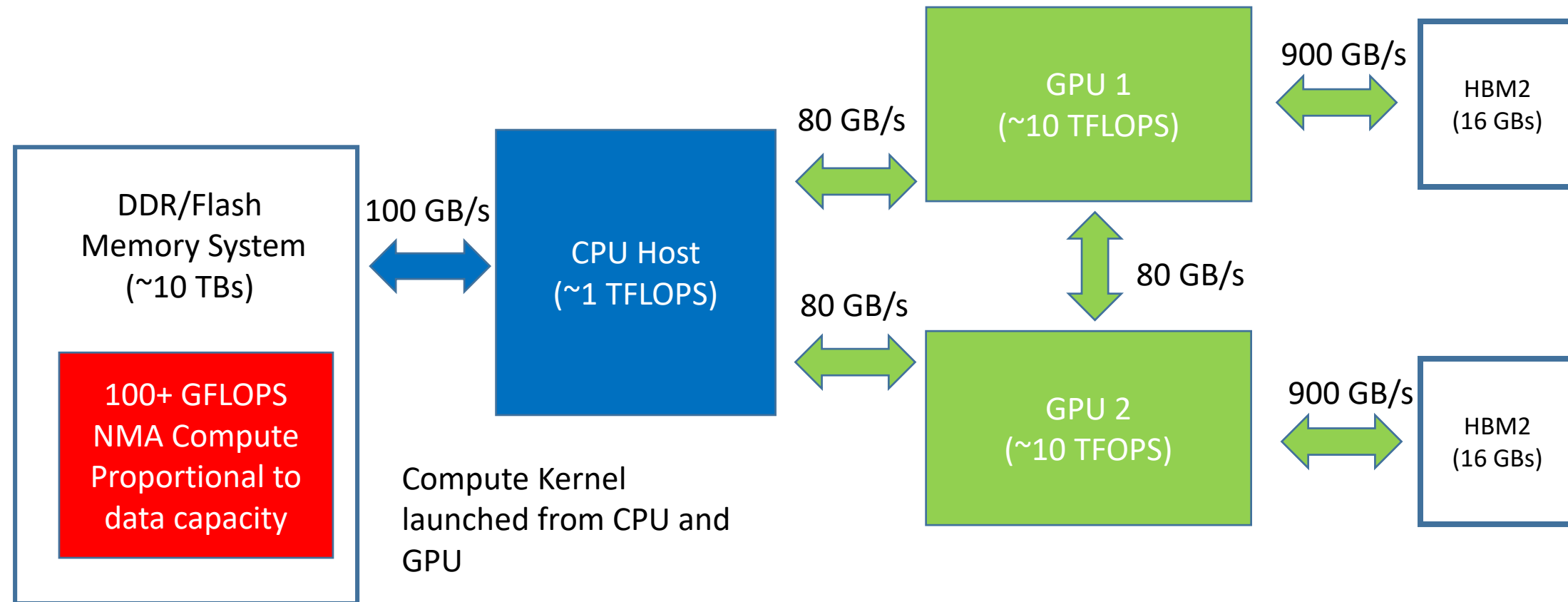


# FlatFlash – Storage-class Memory



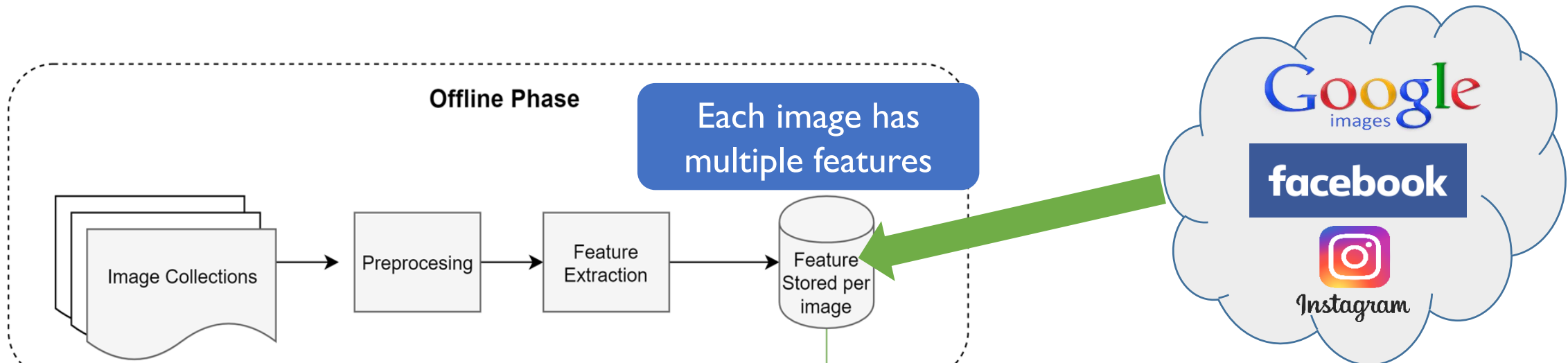
ASPLOS '19 – Abdula, Mailthody, Quresh, Xiong, Huang, Kim, Hwu

# Erudite Vision: place NMA compute inside memory system

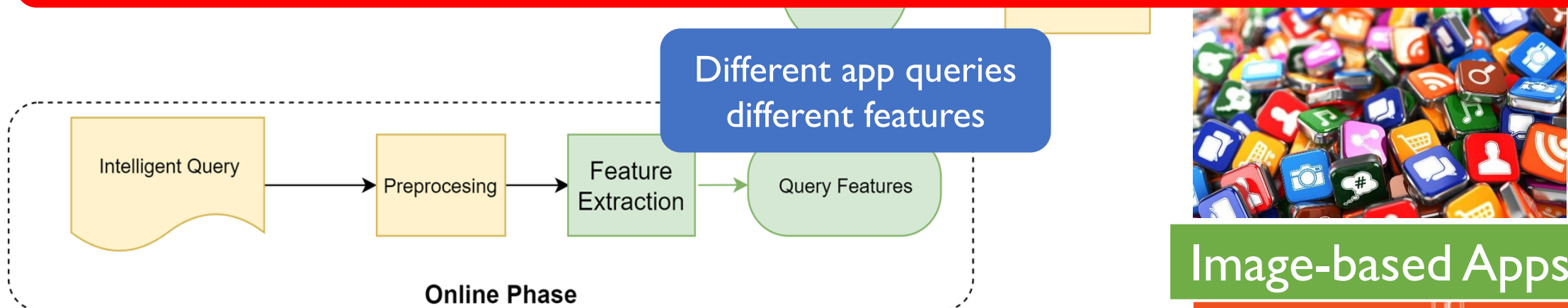


IEEE MICRO 2017

# DeepStore: In-Storage Acceleration for Intelligent Image Search



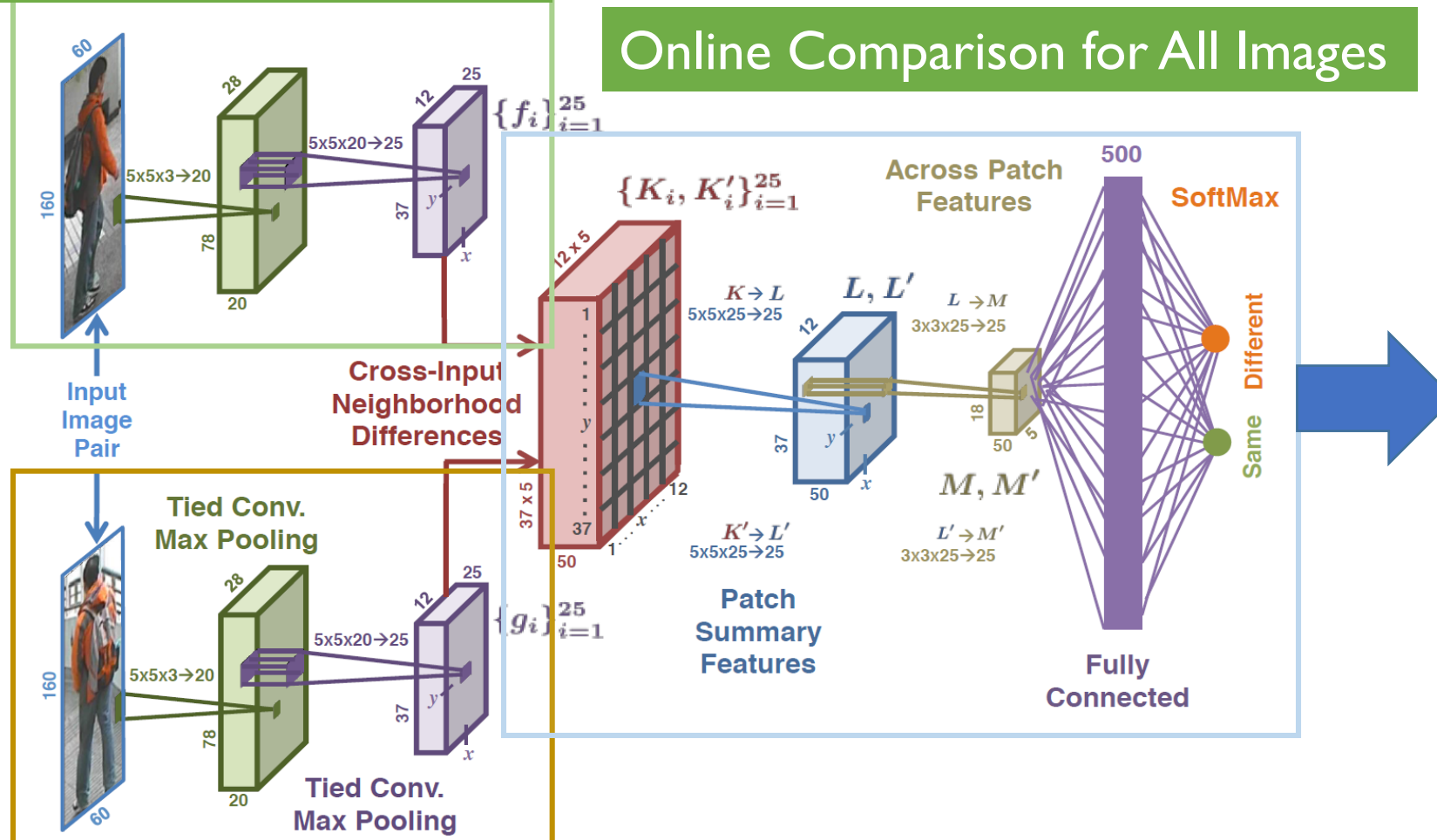
Hard to Build Index for Intelligent Image-Based Search Applications



# Case Study: Person Re-Identification

## Offline Preprocessing

## Online Comparison for All Images



2 convolutions  
1 matrix multiplication  
2 matrix addition  
2 comparison

## Online Query for One Image

# Some Predictions for Yale:100

- Prominent Companies will be very different from today.
- Prominent products will be very different from today.
- The role of universities will be very different from today.
- We will still complain about ISCA and MICRO reviews...
- We will still come to Barcelona.

Way to go, Yale!

