AI 芯片 – AI 芯片基础

计算的工作原理

ZOMI

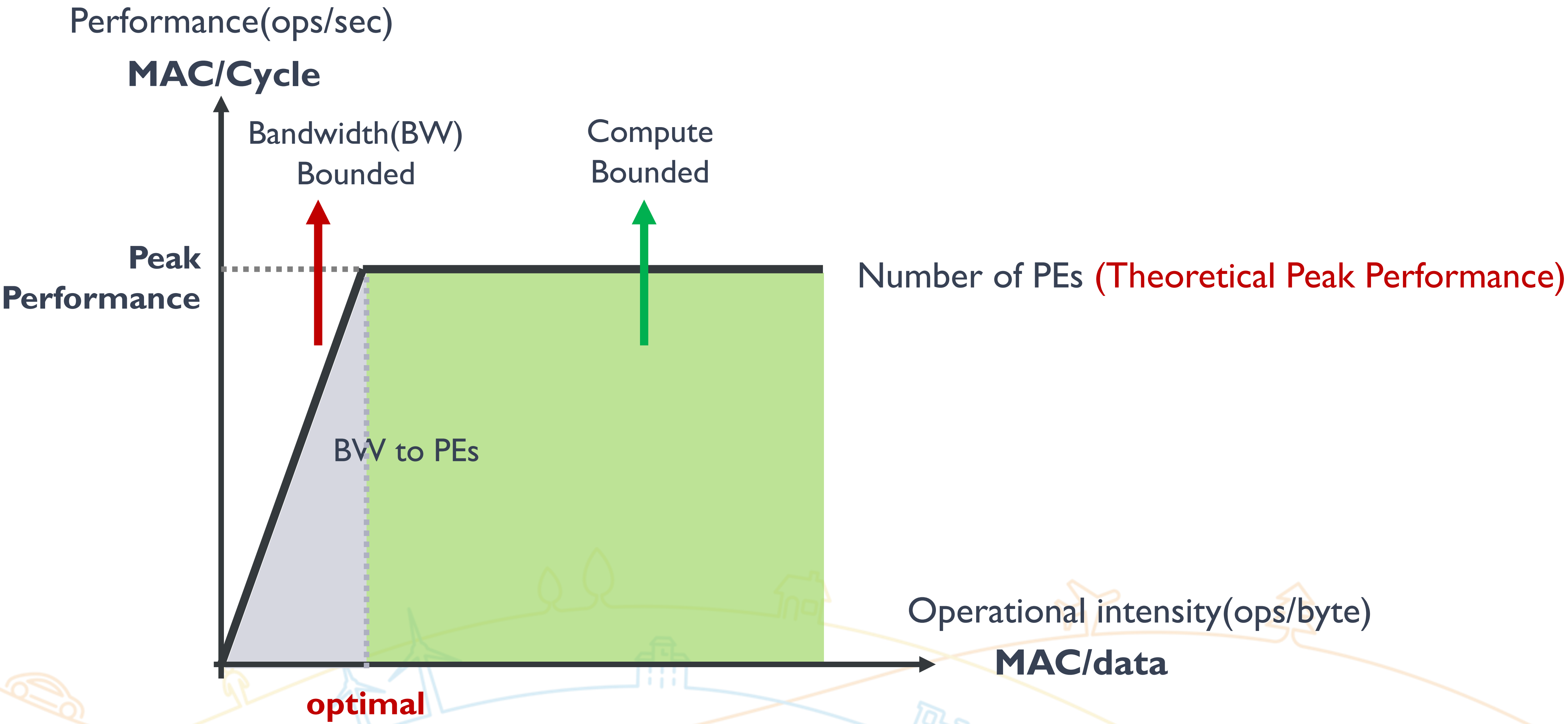
BUILDING A BETTER CONNECTED WORLD

Ascend & MindSpore

www.hiascend.com
www.mindspore.cn

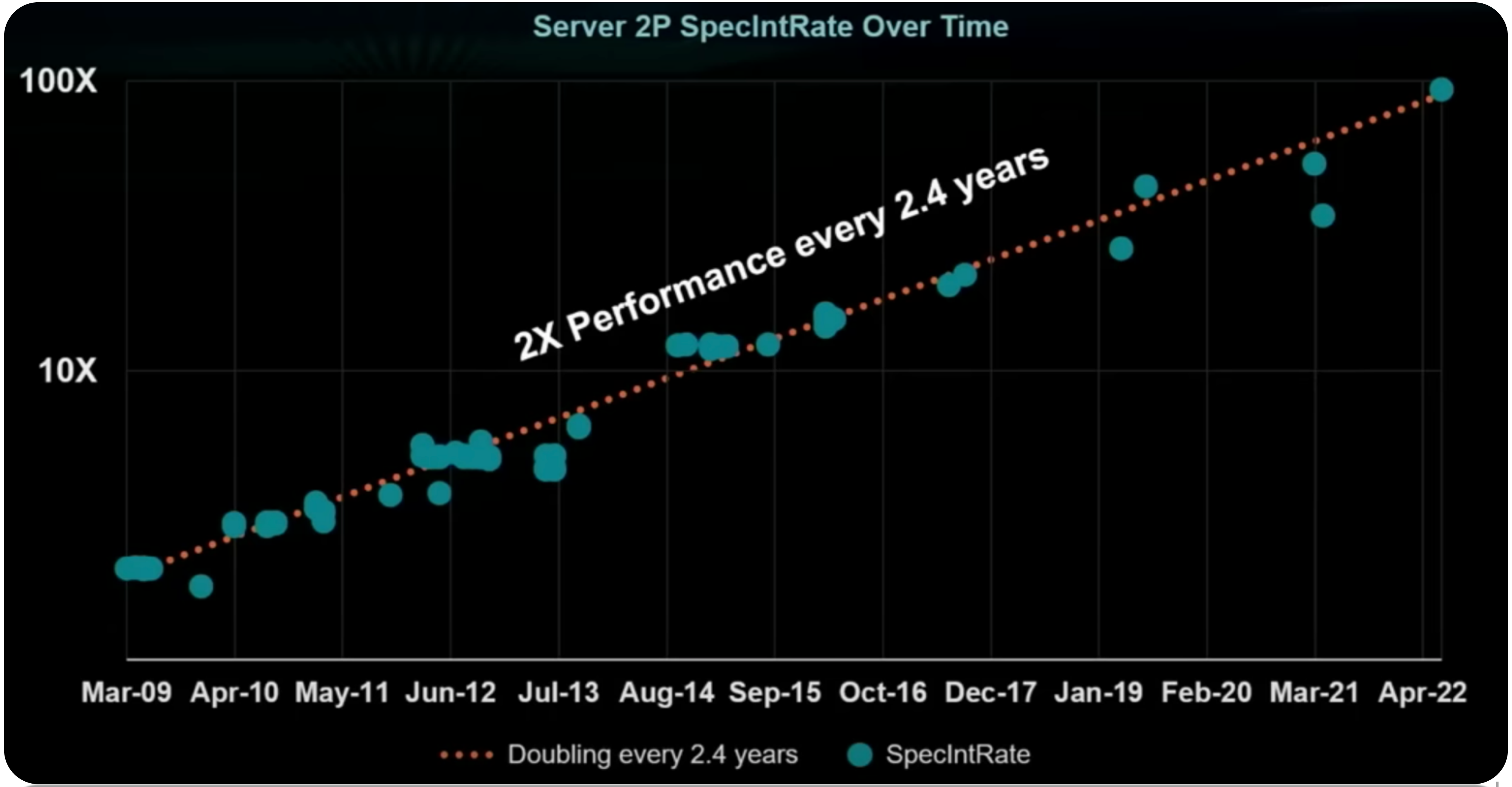
Talk Overview

1. **AI 计算体系**
   - 深度学习计算模式
   - 计算体系与矩阵运算
2. **AI 芯片基础**
   - 通用处理器 CPU
   - 从数据看 CPU 计算
   - 通用图形处理器 GPU
   - AI专用处理器 NPU/TPU
   - 计算体系架构的黄金10年



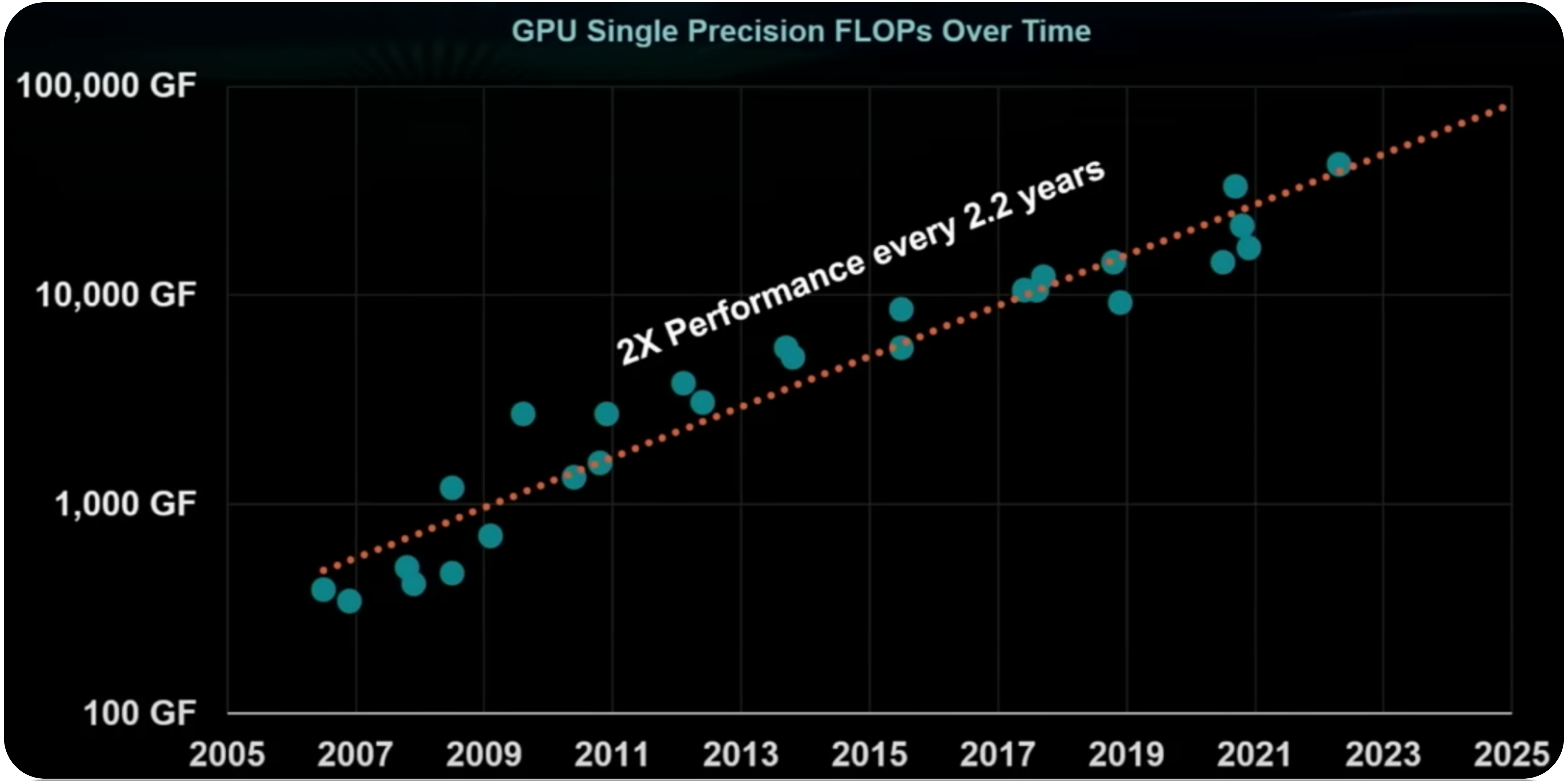
https://arxiv.org/abs/1807.07928

服务器的性能趋势



Server 2P SpecIntRate Over Time
100X
10X
2X Performance every 2.4 years
Mar-09
Apr-10
May-11
Jun-12
Jul-13
Aug-14
Sep-15
Oct-16
Dec-17
Jan-19
Feb-20
Mar-21
Apr-22
Doubling every 2.4 years
SpecIntRate

GPU 性能趋势

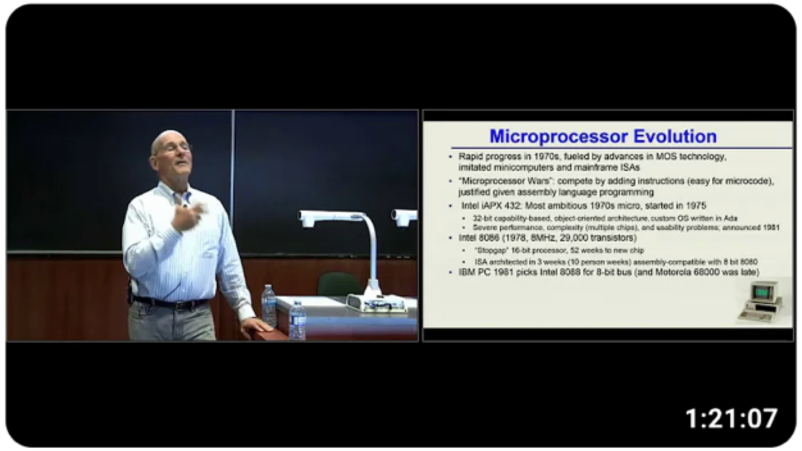


GPU Single Precision FLOPs Over Time
100,000 GF
10,000 GF
1,000 GF
100 GF
2005
2007
2009
2011
2013
2015
2017
2019
2021
2023
2025
2X Performance every 2.2 years

计算机架构的新黄金时代

- A New Golden Age for Computer Architecture: History, Challenges and Opportunities

  https://www.youtube.com/watch?v=kFT54hO1X8M



David Patterson - A New Golden Age for Computer Architecture: History, Challenges and Opportunities

7.1万次观看 • 3年前

UBC Computer Science

Abstract: In the 1980s, Mead and Conway democratized chip design and high-level language programming surpassed assembly ...

Turing Awards | What is Computer Architecture | IBM System360 | Semiconductors | Microprocessor... 44 个章节

编译器的黄金时代

- The Golden Age of Compiler Design in an Era of HW/SW Co-design
- https://www.youtube.com/watch?v=4HgShra-KnY

ASPLOS Keynote: The Golden Age of Compiler Design in an Era of HW/SW Co-design by Dr. Chris Lattner

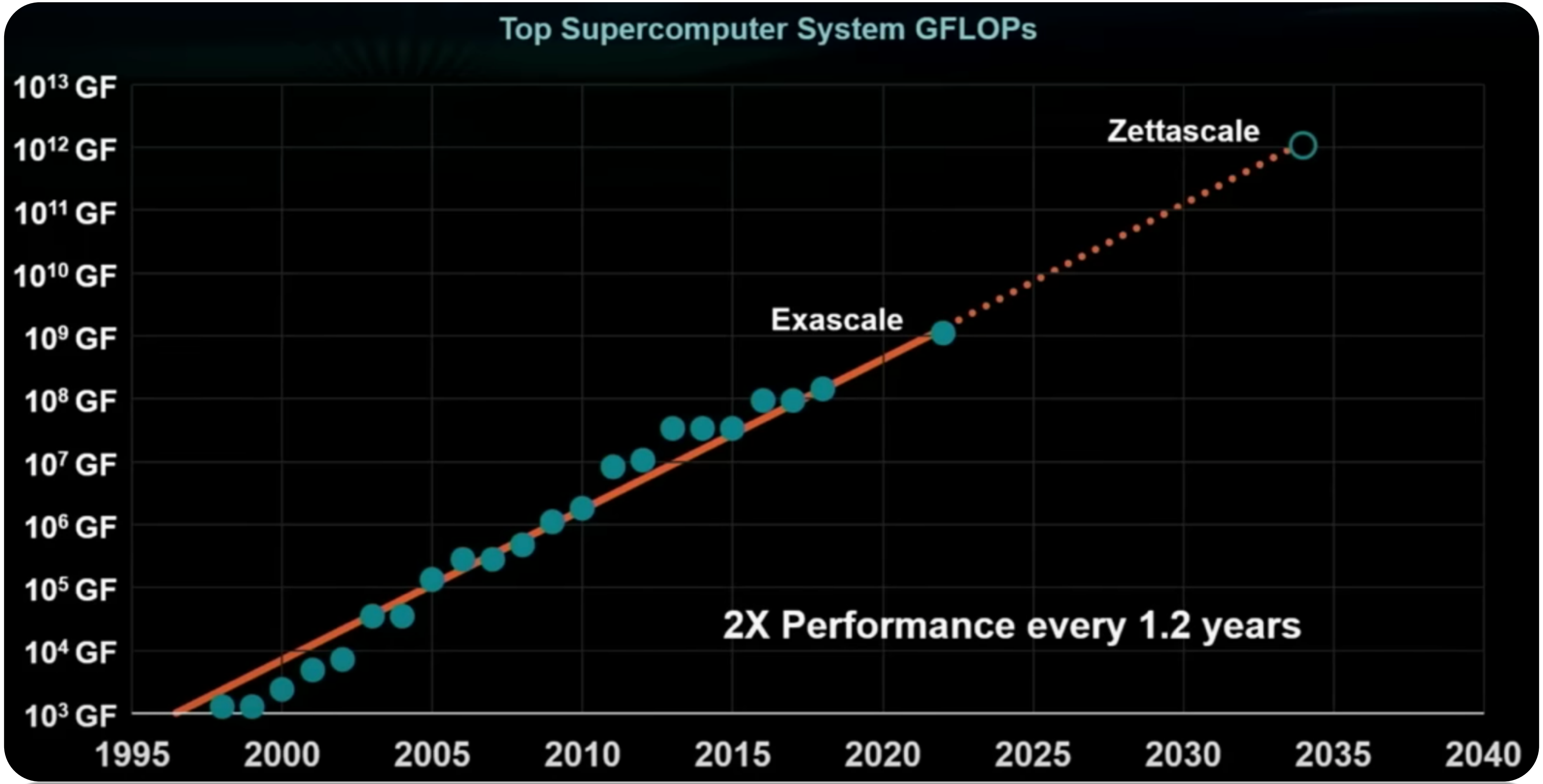
2.7万次观看 • 1年前

SiFiveInc

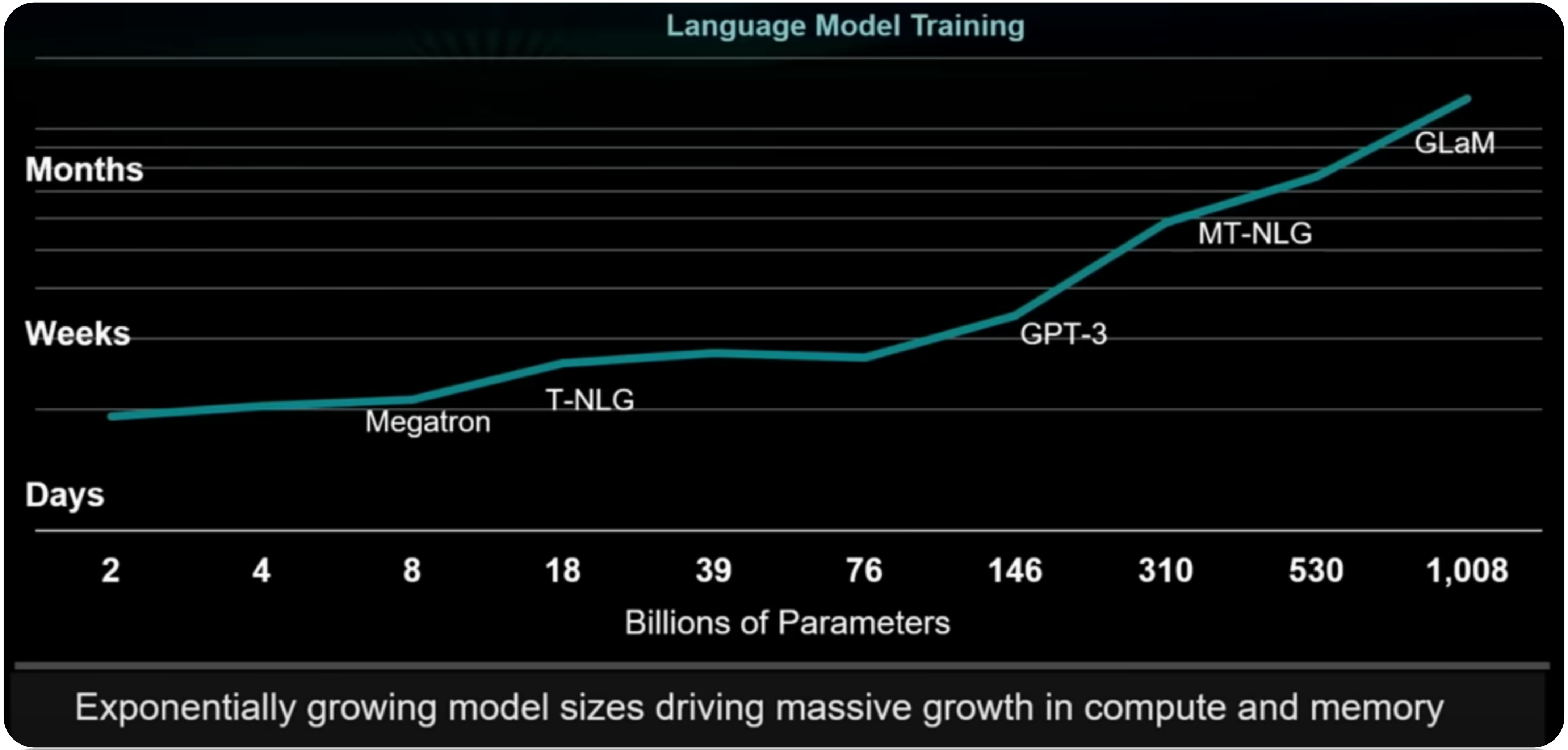
This week at **the** ASPLOS 2021 conference, **Dr. Chris Lattner** gave **the** keynote address to open **the** event with a discussion of **the** ...

A New Golden Age for Computer Architecture John L. Hennessy, David A. Patterson June 2018 End o... 22 个章节

超算中心的性能

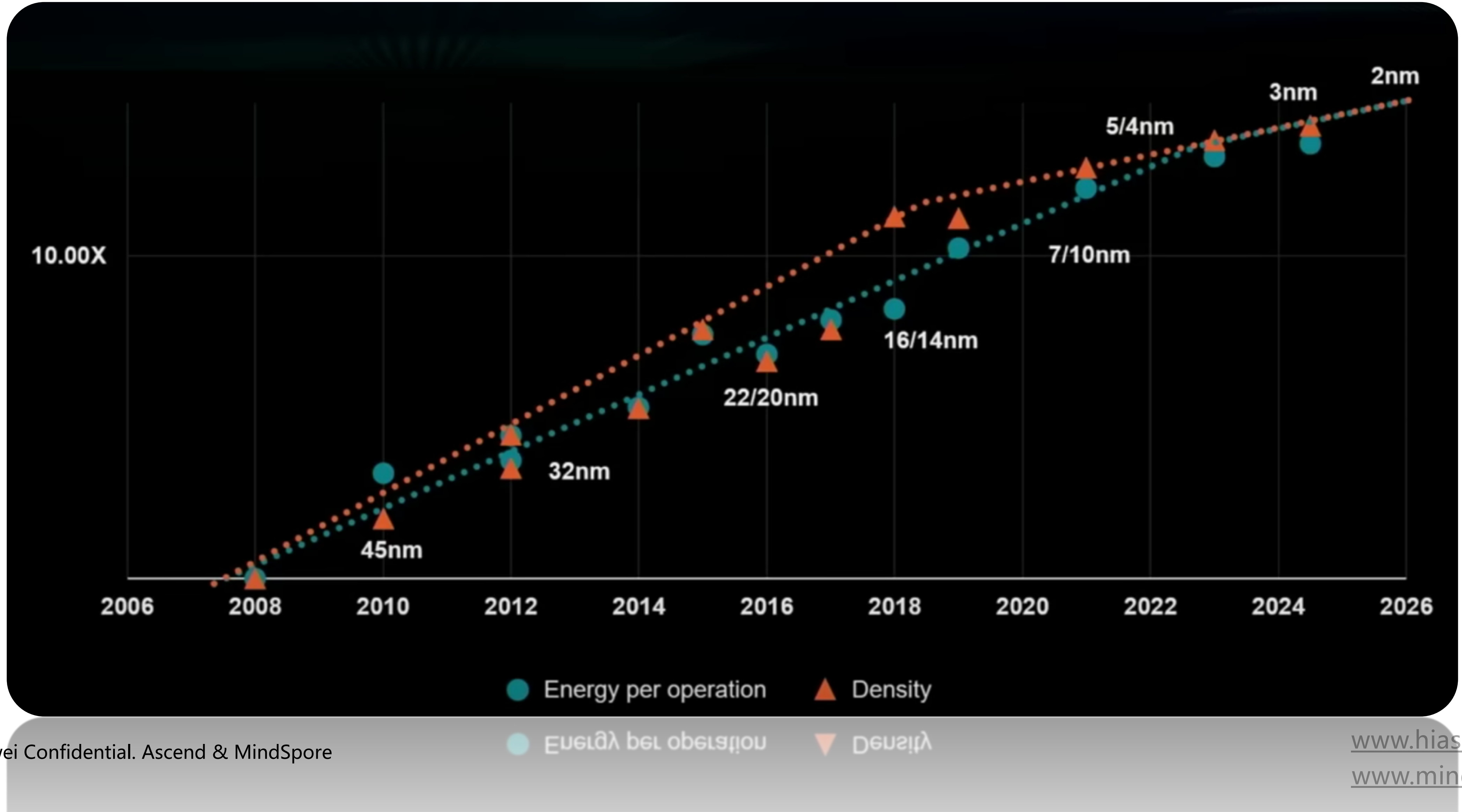

Top Supercomputer System GFLOPs
10^13 GF
10^12 GF
10^11 GF
10^10 GF
10^9 GF
10^8 GF
10^7 GF
10^6 GF
10^5 GF
10^4 GF
10^3 GF
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
Zettascale
Exascale
2X Performance every 1.2 years

训练 AI 大模型的时间



Language Model Training
Months
Weeks
Days
GLaM
MT-NLG
GPT-3
T-NLG
Megatron
2
4
8
18
39
76
146
310
530
1,008
Billions of Parameters
Exponentially growing model sizes driving massive growth in compute and memory

逻辑电路技术趋势预测


2nm
3nm
5/4nm
7/10nm
16/14nm
22/20nm
32nm
45nm
10.00X
2006
2008
2010
2012
2014
2016
2018
2020
2022
2024
2026
Energy per operation
Density

谁会在乎算力呢？

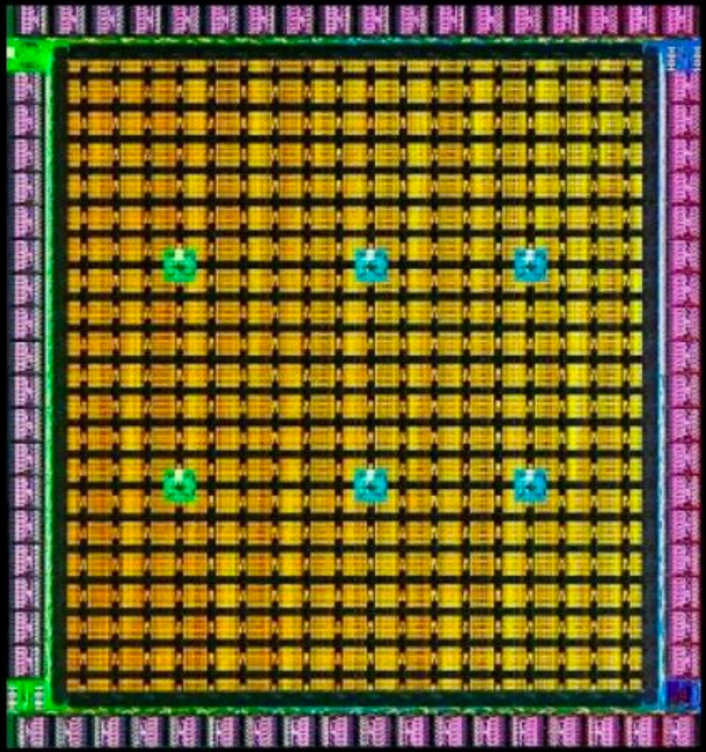
Flops

D1 Chip

362 TFLOPs BF16/CFP8
22.6 TFLOPs FP32

10TBps/dir. On-Chip Bandwidth
4TBps/edge. Off-Chip Bandwidth

400W TDP



645mm²
7nm Technology

50 Billion
Transistors

11+ Miles
Of Wires

BOW: 3RD GENERATION IPU SYSTEMS

SHIPPING TO CUSTOMERS TODAY

BOW POD16

4x Bow-2000
5.6 PetaFLOPS
1 CPU server

BOW POD32

8x Bow-2000
11.2 PetaFLOPS
1 CPU server

BOW POD64

16x Bow-2000
22.4 PetaFLOPS
1-4 CPU server(s)

BOW POD256

64x Bow-2000
89.6 PetaFLOPS
4-16 CPU server(s)

BOW POD1024

256x Bow-2000
358.4 PetaFLOPS
16 - 64 CPU server(s)
Early access

LATEST GPU vs. COLOSSUS Mk2 IPU

	NVIDIA DGX-A100 (8x A100)	GRAPHCORE 8x M2000	
FP32 compute	156TFLOP	2PFLOP	**>12x**
AI compute	2.5PFLOP[1]	8PFLOP[2]	**>3x**
AI Memory	320GB[3]	3.6TB[4]	**>10x**
System Price	$199,000 MSRP	$259,600 MSRP	

NOTES:
[1] Actual figure for TF32/FP16. NVIDIA 8xA100 5PFlop reference is for 50% sparsity which includes Pflops for operations that aren't run
[2] Graphcore AI Float with IEEE FP16.16 multiply.accumulate and IEEE FP16.SR 16bit float with stochastic rounding, with equivalent accuracy performance as FP32
[3] 40GB HBM memory on A100 modules *8 modules per DGX-A100 system
[4] IPU-Exchange Memory which includes attached DRAM and IPU In-Processor-Memory with 100x bandwidth vs. HBM memory sub-system

Flops

你真的在乎算力？

物理定律和硬件本身很大程度决定了我们对机器的编程方式

Flops

你真的在乎算力？

当我们对计算本身有更深入的了解时候，

才会慢慢看到本质的问题：我的数据在哪里？

读取数据与计算的计算换算

How many operations must I do on some data to make it worth the cost of loading it?

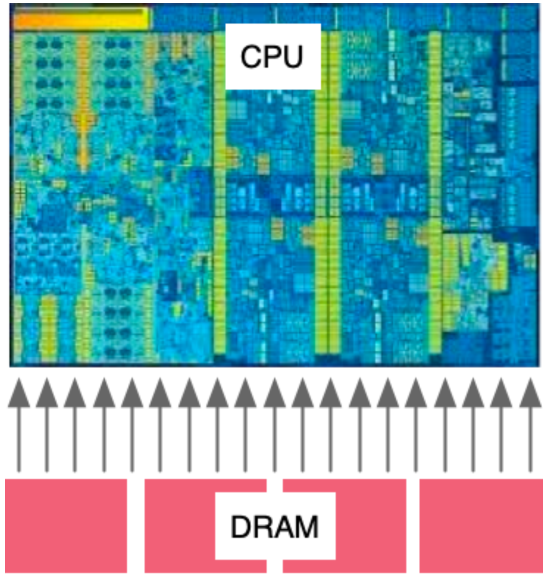
Required Compute Intensity = $\frac{FLOPs}{Data\ Rate}$ = 80

2000 GFLOPs FP64

200 GBytes / sec

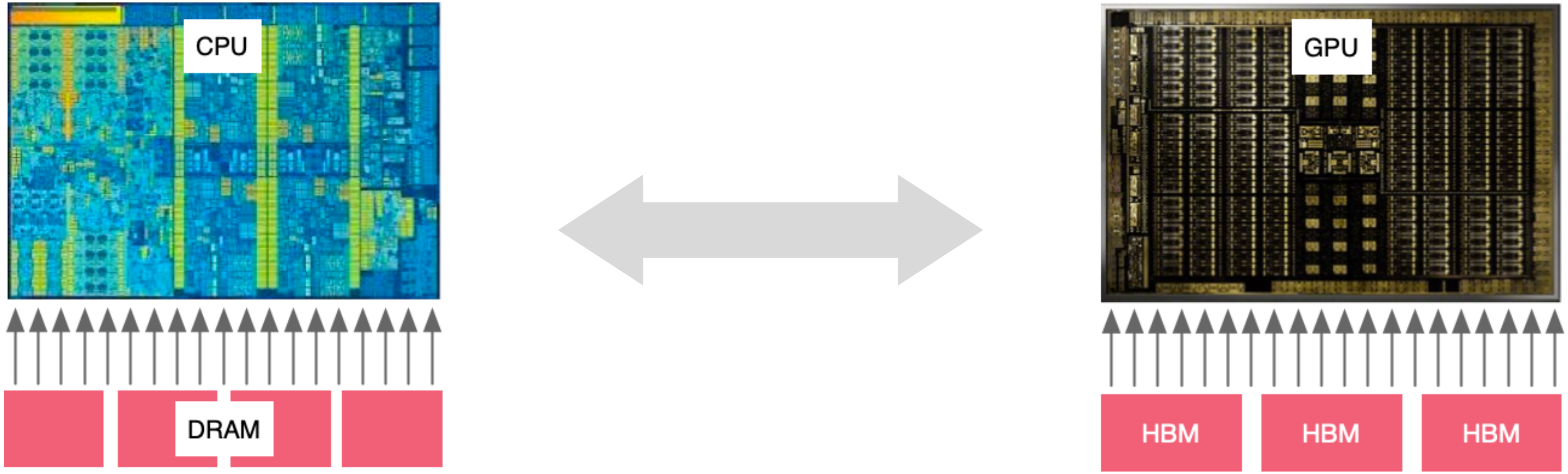
= 25 Giga-FP64 / sec

(FP64 = 8 bytes)



So for every number load from memory, Need to do 80 Operations on it to break even.

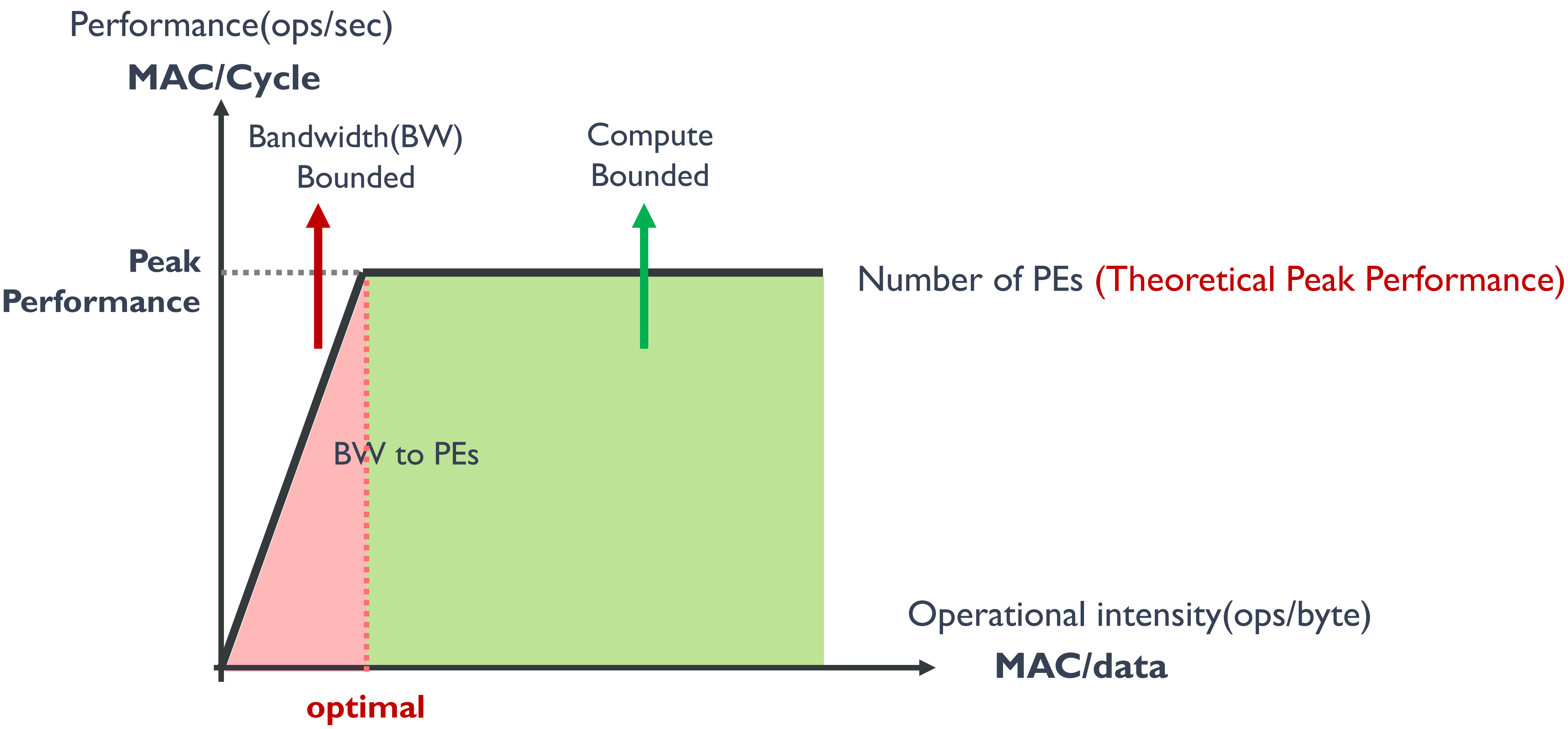
计算密集型



	AMD Rome 7742	Intel Xeon 8280	NVIDIA A100
Peak FP64 Giga Flops	2,190	2,300	19,500
Memory B/W (GB/sec)	131	204	1,555
Compute Intensity	134	90	100

ZOMI并不是很在乎

Flops 算力



https://arxiv.org/abs/1807.07928

更应该关注

内存、带宽、时延

引用

1. https://www.youtube.com/watch?v=3jHi8E5C-18
2. https://www.youtube.com/watch?v=-P28LKWTzrI
3. https://www.youtube.com/watch?v=3l10o0DYJXg

BUILDING A BETTER CONNECTED WORLD

THANK YOU