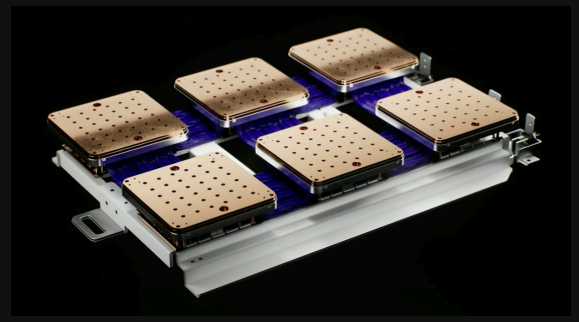
Improving AI Infrastructure Efficiency with Optical I/O

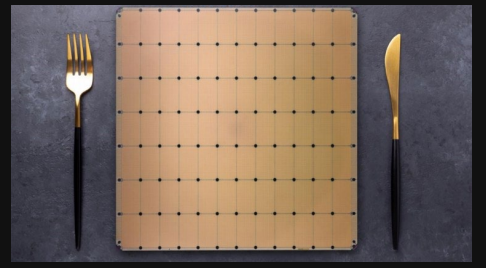
Miloš Popović | Senior Technical Architect & Co-Founder
milos@ayarlabs.com

Optical I/O transformation of compute system scaling

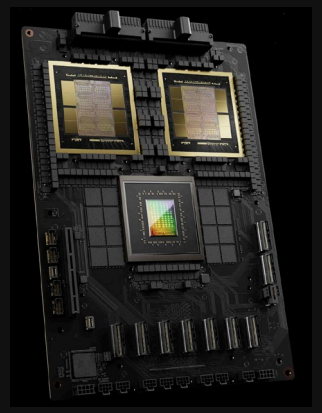


Tesla Dojo (Tesla AI Day 2022).

Cerebras CS-2 wafer size chip.

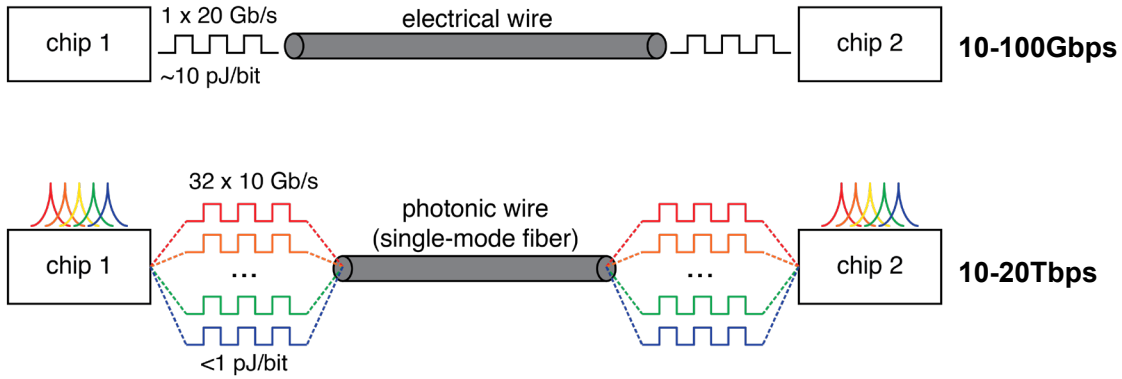


Nvidia GTC 2024 Keynote.

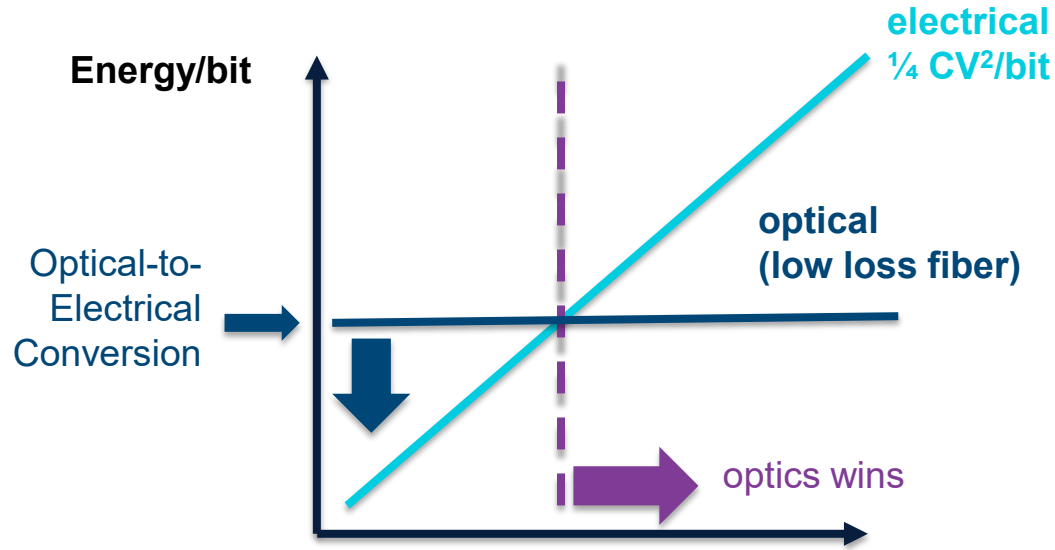


AI compute architectures limited by data movement:

- I/O communication **energy** and **bandwidth density**



Optics provides bandwidth density: many wavelengths, smaller.

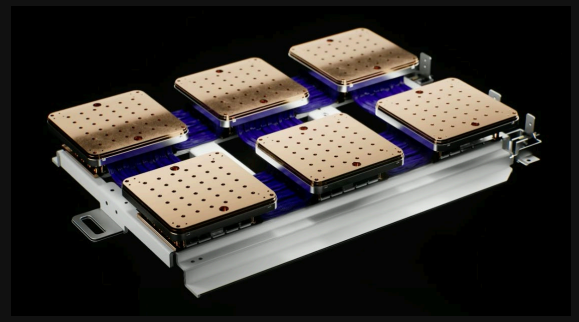


DAB Miller, Opt. Lett. (1990).

Optics provides low energy – requires deep integration.

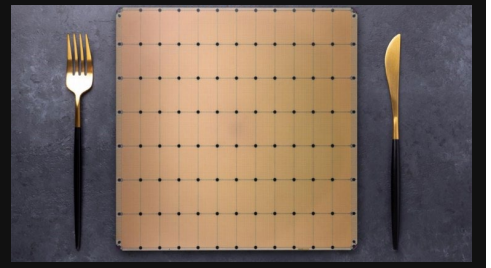
Miloš Popović, Ayar Labs (milos@ayarlabs.com)

Optical I/O transformation of compute system scaling

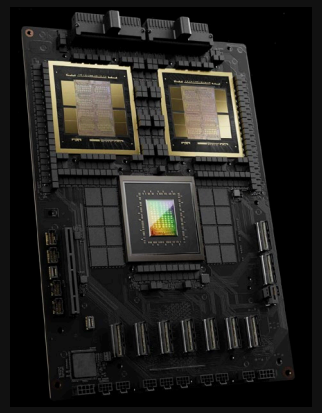


Tesla Dojo (Tesla AI Day 2022).

Cerebras CS-2 wafer size chip.

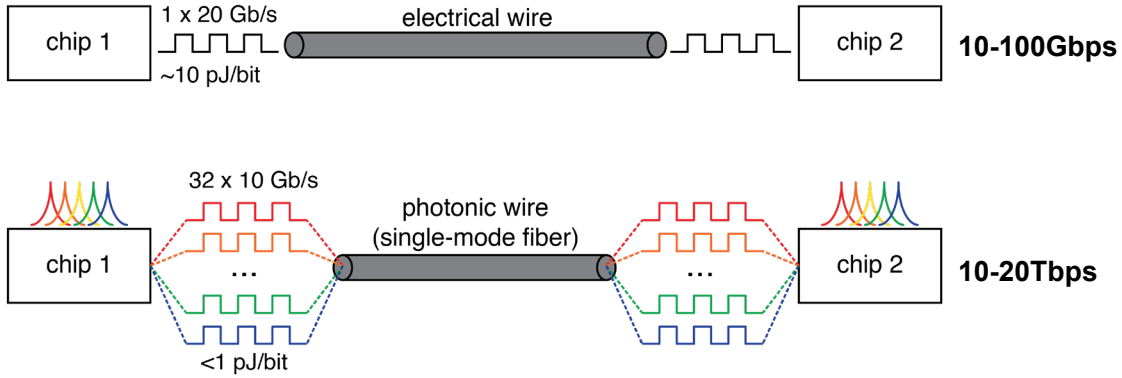


Nvidia GTC 2024 Keynote.

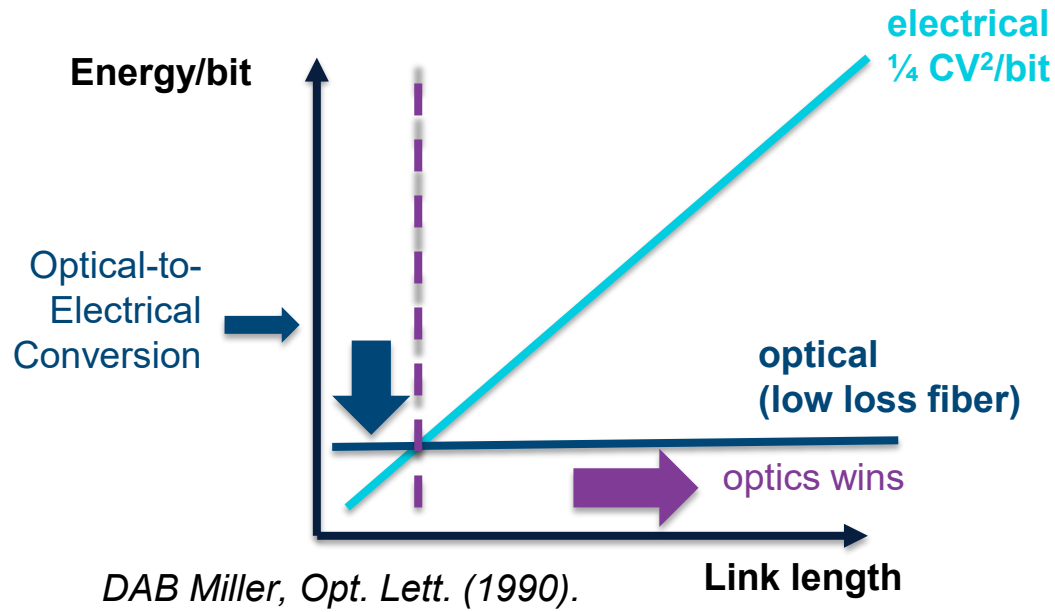


AI compute architectures limited by data movement:

- I/O communication **energy** and **bandwidth density**



Optics provides bandwidth density: many wavelengths, smaller.



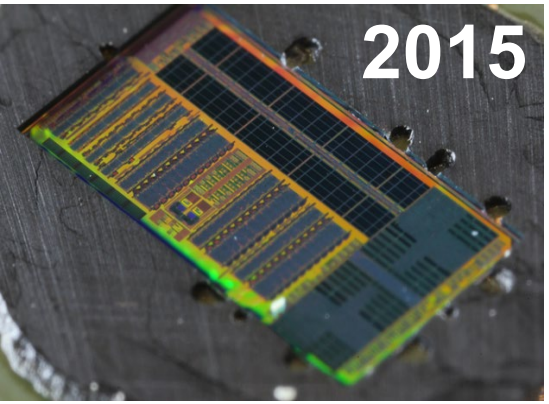
DAB Miller, Opt. Lett. (1990).

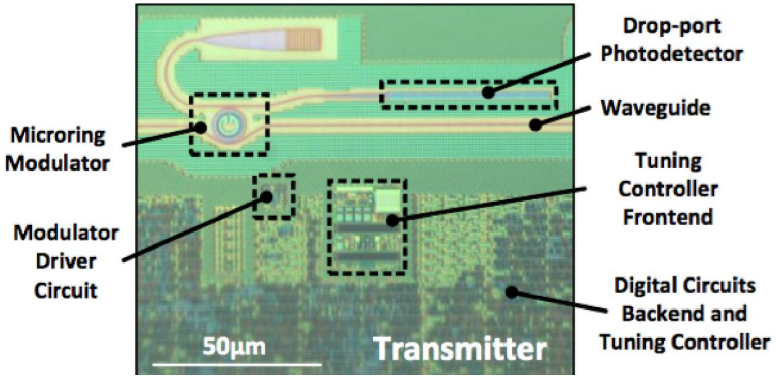
Optics provides low energy – requires deep integration.

Miloš Popović, Ayar Labs (milos@ayarlabs.com)

Optical I/O chiplet enabled interconnection of compute

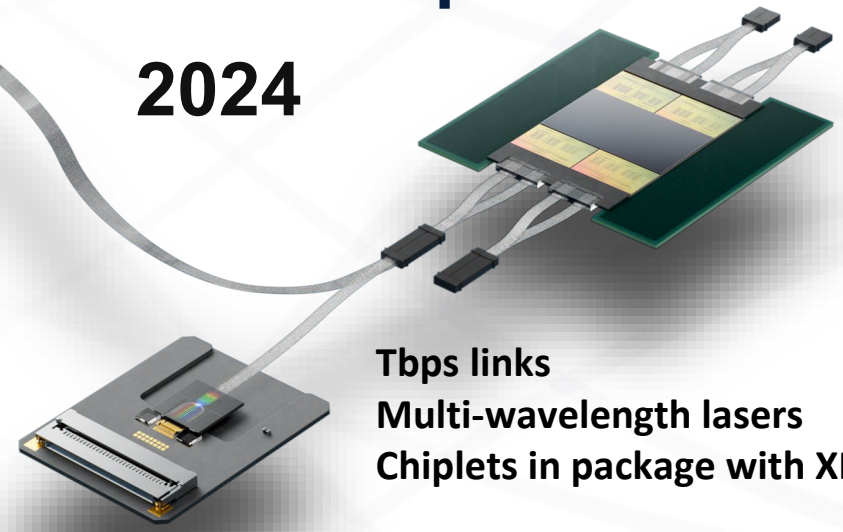
1





2

2024

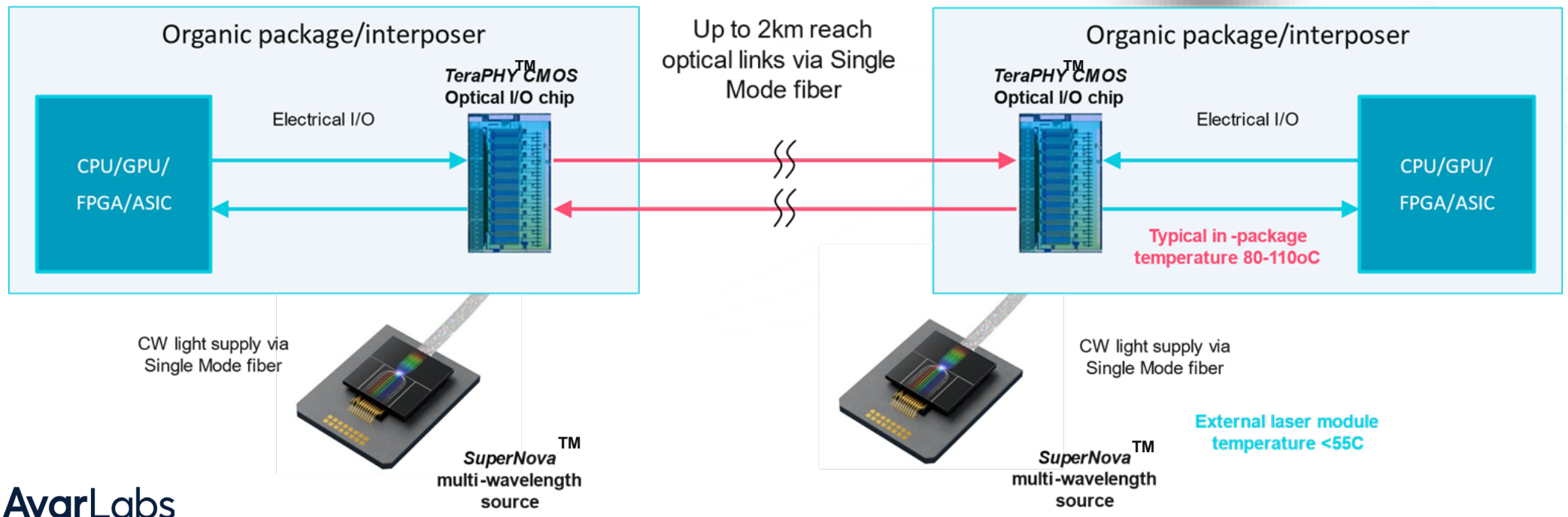


Tbps links
Multi-wavelength lasers
Chiplets in package with XPU

- First microprocessor with light based interconnect (2015)
- University research -> commercial 300mm CMOS fab

Berkeley UNIVERSITY OF CALIFORNIA · BOSTON UNIVERSITY · University of Colorado Boulder · MIT

3



AyarLabs

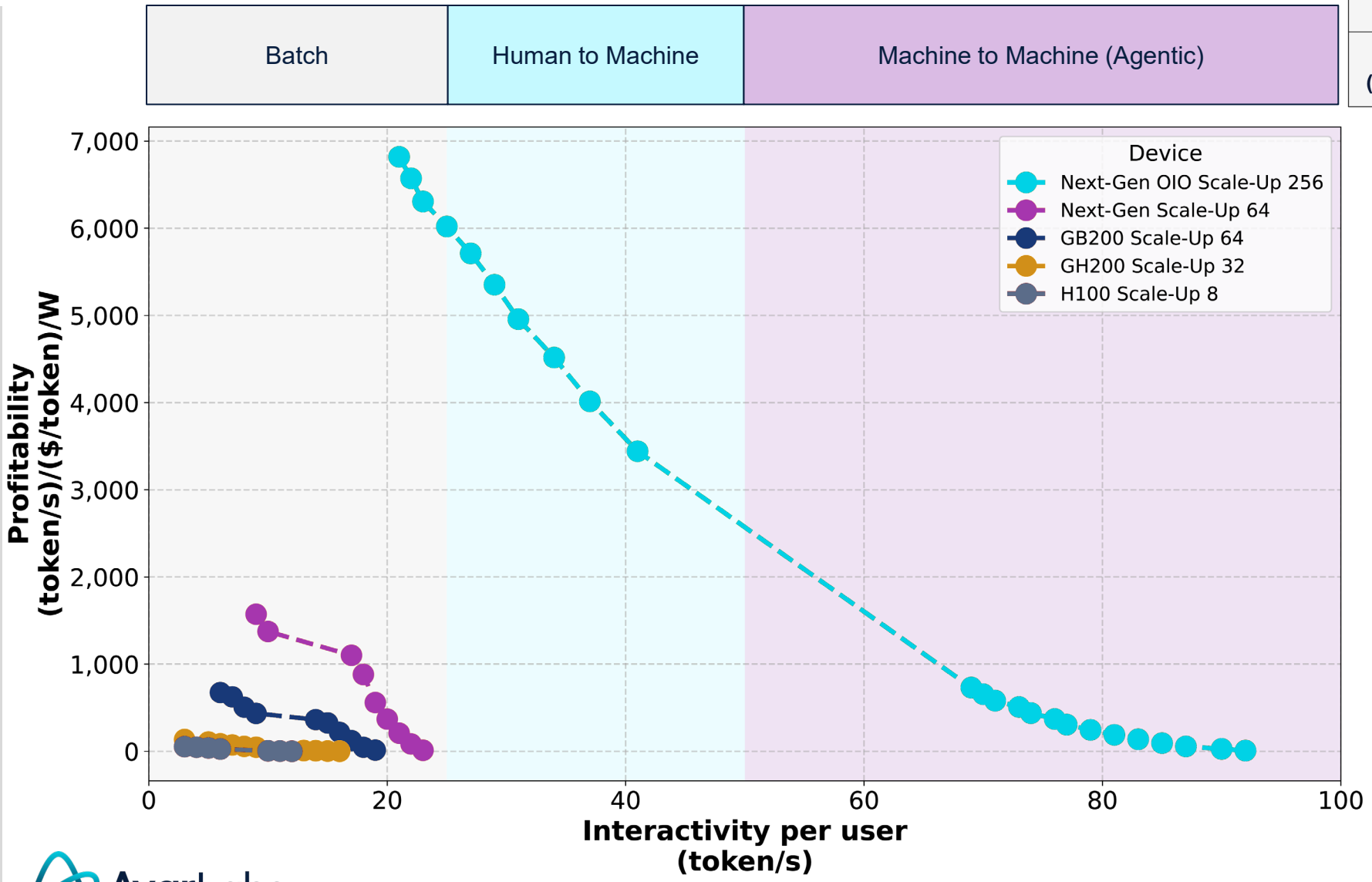
Wade et al., "An Error-free 1 Tbps WDM Optical I/O …", OFC postdeadline (2021).

Miloš Popović, Ayar Labs (milos@ayarlabs.com)

Optical I/O impact on future GPT-X performance

Figure of Merit	Description	Units
Throughput	$\frac{\text{Number of Users}}{\text{Time Per Output Token}}$	$\frac{\text{Token}}{\text{Second}}$
Interactivity	$\frac{1}{\text{Time Per Output Token}}$	$\frac{\text{Token}}{\text{Second}}$
Profitability (AI efficiency)	$\frac{\text{Throughput}}{\text{Cost X Power}}$	$\frac{\text{Token / Second}}{\frac{\$}{\text{Token}} \text{ X Watts}}$

AI efficiency (performance / $ / W)



State of Art ~2026-27

14T parameters
Mixture of Experts (MoE) w/ 32 experts

inference with:
32K/8K sequence length
5s Time to First Token

Optimizations:

Tensor/Sequence parallel
Pipeline parallel
Expert parallel

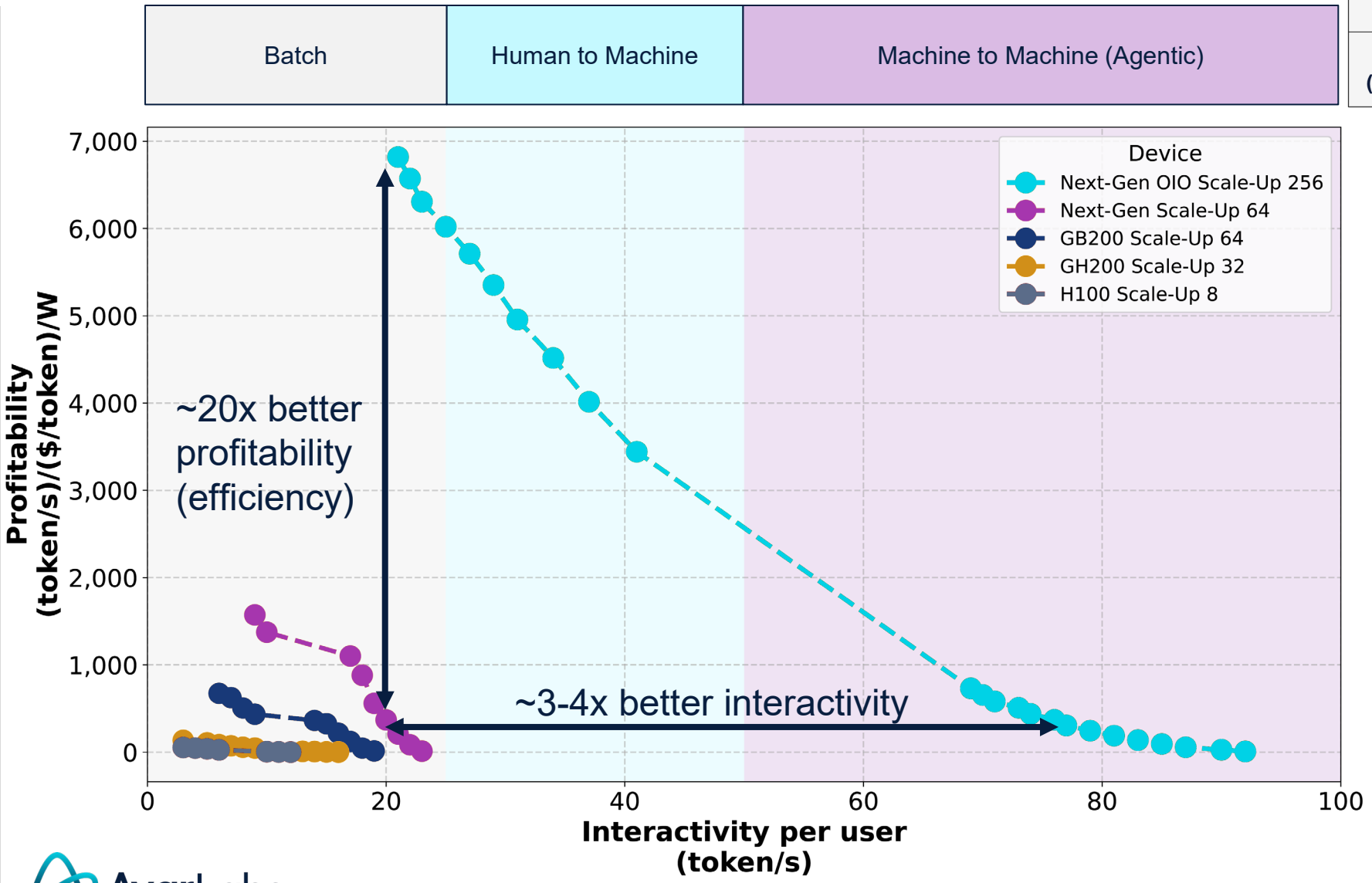
AyarLabs

Source: Ayar Labs CTO office team (Oct 2024)

Optical I/O impact on future GPT-X performance

Figure of Merit	Description	Units
Throughput	$\frac{\text{Number of Users}}{\text{Time Per Output Token}}$	$\frac{\text{Token}}{\text{Second}}$
Interactivity	$\frac{1}{\text{Time Per Output Token}}$	$\frac{\text{Token}}{\text{Second}}$
Profitability (AI efficiency)	$\frac{\text{Throughput}}{\text{Cost X Power}}$	$\frac{\text{Token / Second}}{\frac{\$}{\text{Token}} \text{ X Watts}}$

AI efficiency (performance / $ / W)



State of Art ~2026-27

14T parameters
Mixture of Experts (MoE) w/ 32 experts

inference with:
32K/8K sequence length
5s Time to First Token

Optimizations:

Tensor/Sequence parallel
Pipeline parallel
Expert parallel

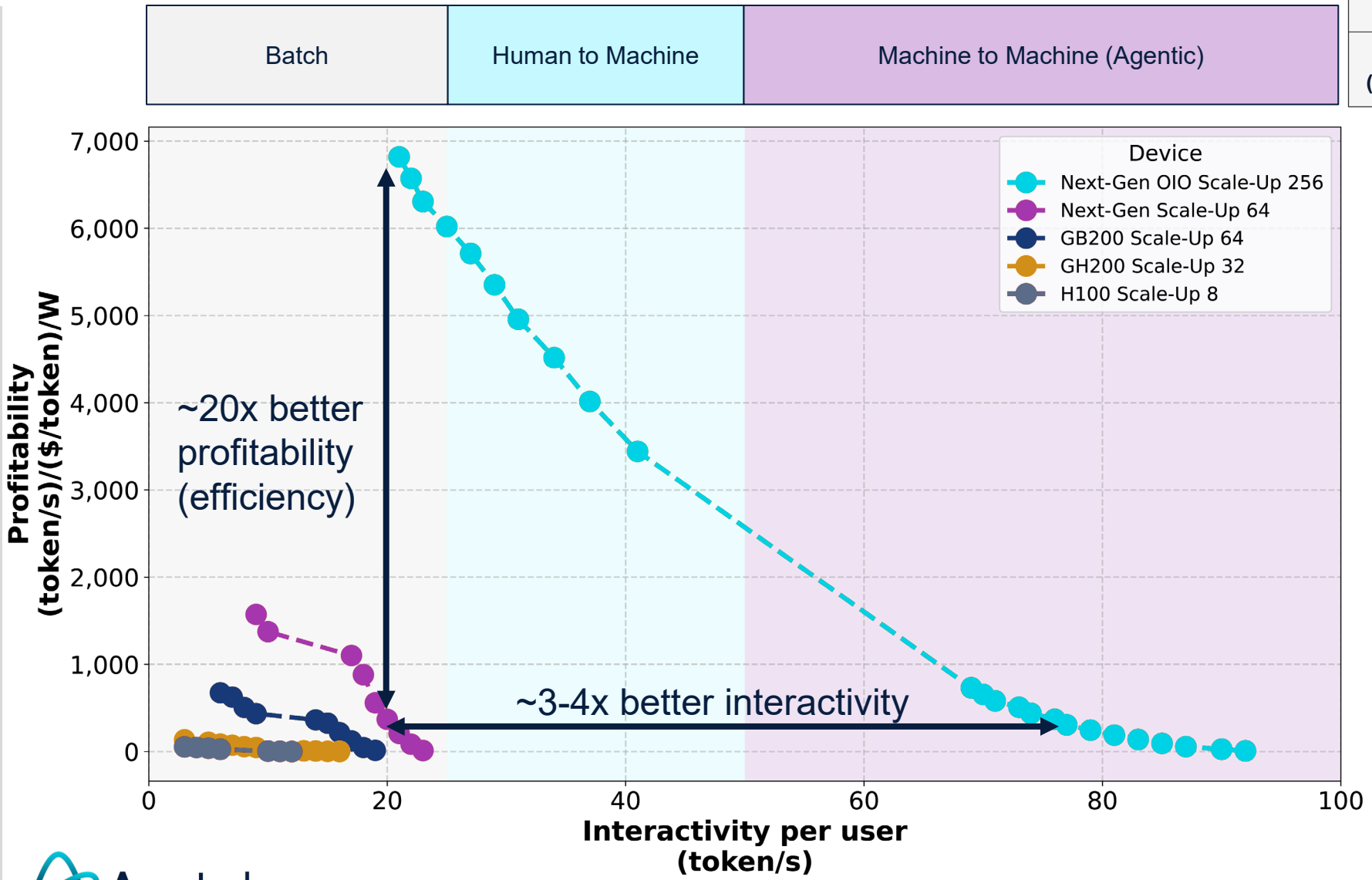
AyarLabs

Source: Ayar Labs CTO office team (Oct 2024)

Optical I/O impact on future GPT-X performance

Figure of Merit	Description	Units
Throughput	$\frac{\text{Number of Users}}{\text{Time Per Output Token}}$	$\frac{\text{Token}}{\text{Second}}$
Interactivity	$\frac{1}{\text{Time Per Output Token}}$	$\frac{\text{Token}}{\text{Second}}$
Profitability (AI efficiency)	$\frac{\text{Throughput}}{\text{Cost X Power}}$	$\frac{\text{Token / Second}}{\frac{\$}{\text{Token}} \text{ X Watts}}$

AI efficiency (performance / $ / W)



State of Art ~2026-27

14T parameters
Mixture of Experts (MoE) w/ 32 experts

inference with:
32K/8K sequence length
5s Time to First Token

Optimizations:

Tensor/Sequence parallel
Pipeline parallel
Expert parallel

AyarLabs

Optical I/O is an opportunity for major near-term AI hardware efficiency impact.

Source: Ayar Labs CTO office team (Oct 2024)