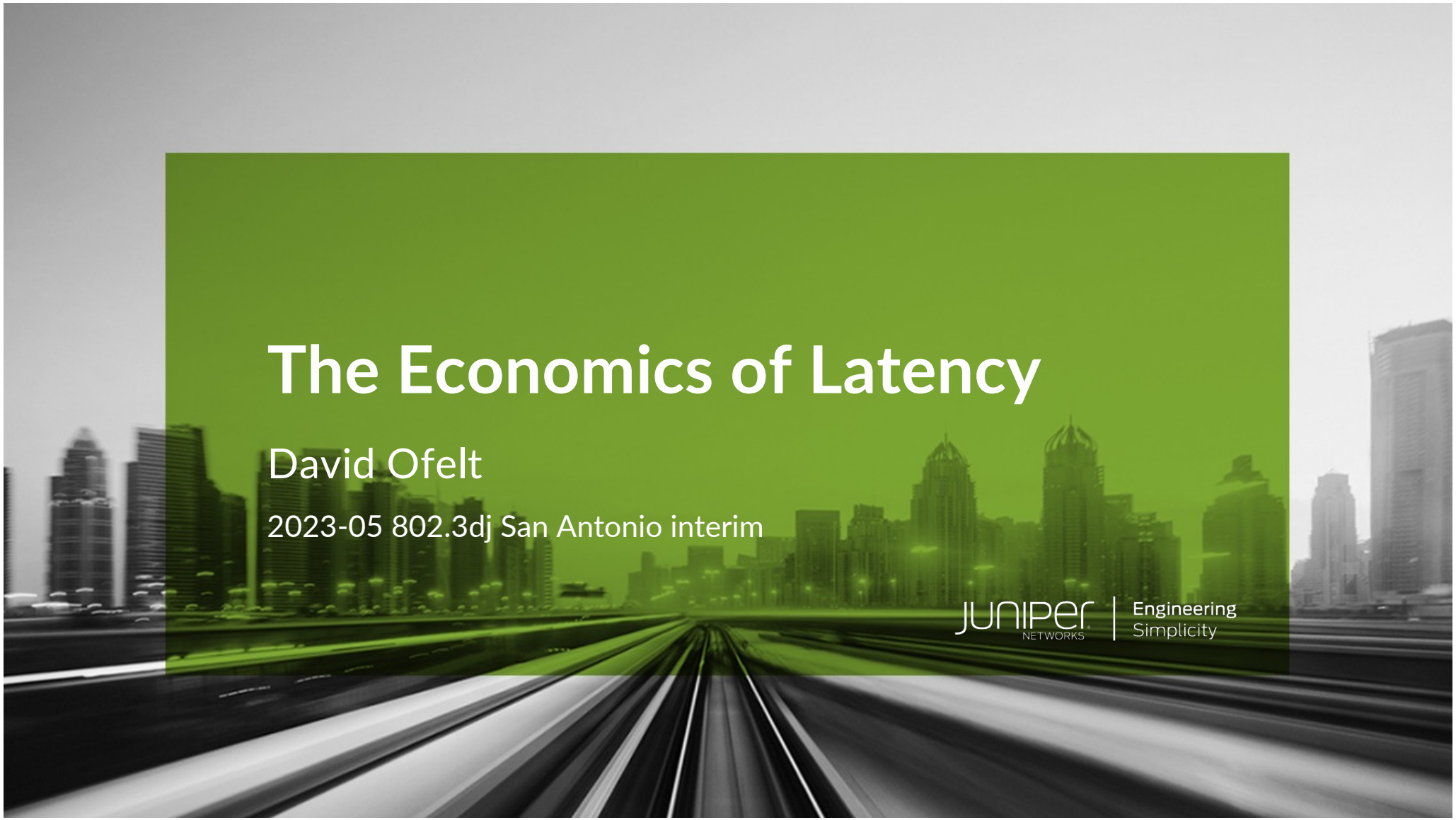
The Economics of Latency
David Ofelt
2023-05 802.3dj San Antonio interim
JUNIPER NETWORKS
Engineering Simplicity

Latency is Important!

Latency is Important!

But….

Is it *more* or *less* important than <*your favorite thing*>?

Tradeoffs

802.3 is very used to tradeoffs… in every project we find balance between a large number of topics

- reach
- BER
- area
- power
- composability
- complexity
- cost
- manufacturability
- technology scaling
- technology evolution
- implementation diversity
- reuse
- latency
- etc, etc

These tradeoffs are done with the goal of maximizing the Broad Market Potential

We have internalized the architectures used in datacenters, metro networks, carrier networks, etc

- Many presentations over the years on concrete details and tradeoffs

Why has interest in lower latency increased?

We have always made some attempt to optimize latency, but….

Recently the explosion of the AI/ML market has brought more focus on the topic

(My opinion) AI/ML is really just classic High Performance Computing (HPC) with better marketing

- HPC == Heavy Matrix Math
- AI/ML difference is that it sometimes add new, weird, floating-point formats

HPC traditionally is sensitive to latency

- many algorithms are sensitive to the ratio between computation and communication
- sensitivity is deeply dependent on the details of the algorithms
- there is great diversity in algorithms used
- there is great diversity in architectures being used

There is no single ”AI/ML” market, architecture, or application

- Just like there is no single Ethernet application

Questions for industry…

We need to know more about the details of the AI/ML/HPC architectures, applications, and market

- Then them to the rest of our tradeoffs

Showing up and just stating "I need lower latency **because AI/ML !!1!!!**" is not enough ☺

I'd like to know things like: For an X% increase in latency- what is the % increase in … ?

- runtime
- compute resources
- accuracy
- etc

What is the full end-to-end latency and what part of that is in 802.3's control?

- Seemingly dramatic differences at the PCS/PMD level may be not be significant at the CPU core to CPU core level

The speed of light is slow

- given that- what reaches are really latency sensitive?

More questions…

We need to know more about the market….

How many of the links that are being called “AI/ML” links are really latency sensitive?

How many of the links are ”inside the machine” or connecting machines?

Are the links actually using Ethernet or are they repurposing Ethernet technology?

For a given PMD- how many of the latency sensitive links -vs- the rest of the use cases for that PMD?

- If we add cost and power or limit the technology choices to improve latency- what does that do to the broad market potential

Ethernet Just Works

We try to make our PMDs Just Work by default

- We attempt to make reliability high by default
- Attempt to make interop be as easy as possible

We provide advanced knobs for experts

- Engineered links can get more reach or rely on a weaker (but lower latency) FEC
- Guard-rails that add latency can be bypassed (ex: FEC indication bypass)
- This is an excellent approach

Non-Juniper

Closing thoughts

A large amount of what we do in 802.3 is maximizing broad market potential

Market Changes

- More and more of HPC is moving to Ethernet
- AI/ML is an exciting new part of the HPC market that is growing rapidly
- These benefit from lower latency

Lower latency is important, but how do we trade it off against everything else?

The need for lower latency is not going to go away, so as a group, we need to figure out the tradeoffs

Providing expert knobs to allow for lower latency can work well

Thanks!