Latency for Dummies

The bane of our existence

Friday, July 31

Overview of Mirazon

- Consulting services company
- Founded in 2000
- Focus on partnering with companies to facilitate success
- T&M or “managed” agreements – we offer both and don’t require retainers
- Offer best-of-breed technology solutions that are rigorously tested and that we are highly certified and trained
- Best Place to Work in Greater Louisville

AND Kentucky five years running

Let Me Introduce Myself

- Worked at Mirazon since 2007
- Currently Chief Technology Officer
- MCSE 2003, 2016 (different acronyms), MCITP-EA
- SME for Microsoft for Hyper-V 2008
- VMware VCP 3.5, 4, 5, 6, 6.5, VCAP DCA: 4, 5, DCD 5
- BS:CIS Information Security - UofL

Disclaimer

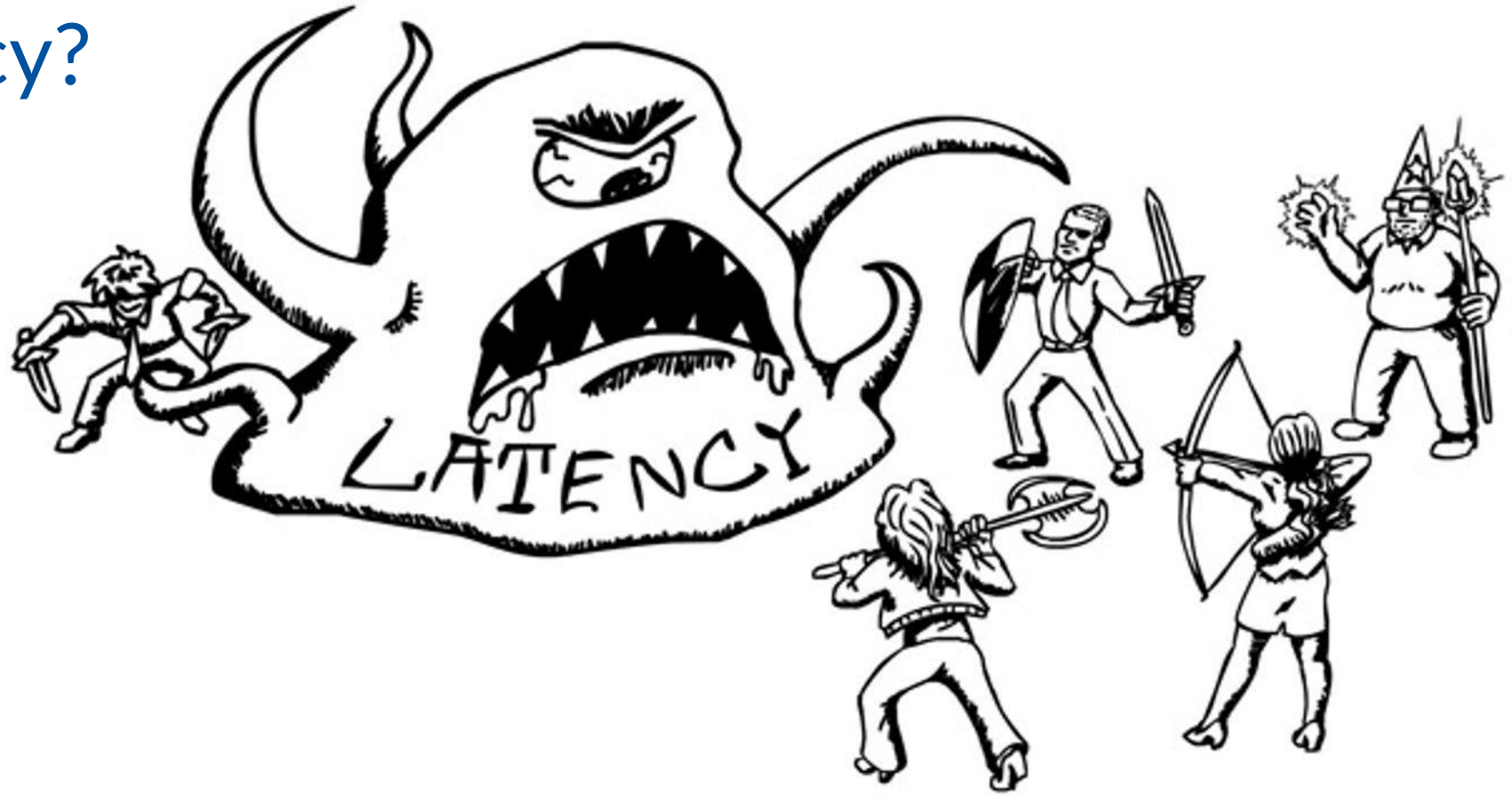
I'm not an electrical engineer, physicist, mathematician, or any of the other over-educated people who truly understand the complex and variable-rich math involved with some of the calculations referenced in this presentation.

Some of these numbers took hours of research and reading far more about physics than I cared to, simply to get something as straightforward as "the industry states that a good rule of thumb is .66."

Every number in this presentation has been rounded to an easy-to-discuss number, so don't throw things if you don't agree with a number specifically.

Agenda

- What is latency?
- Interconnects
- Network
- Server
- Storage
- THE CLOUD!
- Users



Latency

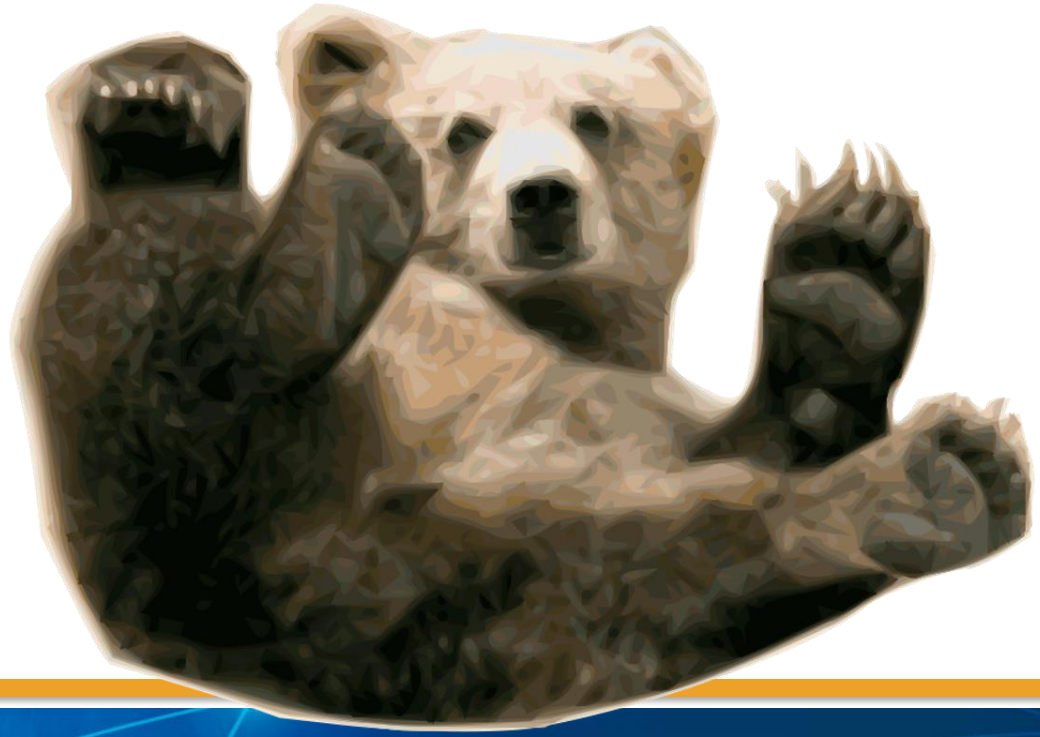
What is it?

Dictionary.com:

“The period of apparent inactivity between the time the stimulus is presented and the moment a response occurs.”

Translated:

The length of time between when you poke the bear and when it kills you.”



Latency

How is it measured?

1 second =

1,000 milliseconds (ms)

1,000,000 microseconds (µs)

1,000,000,000 nanoseconds (ns)

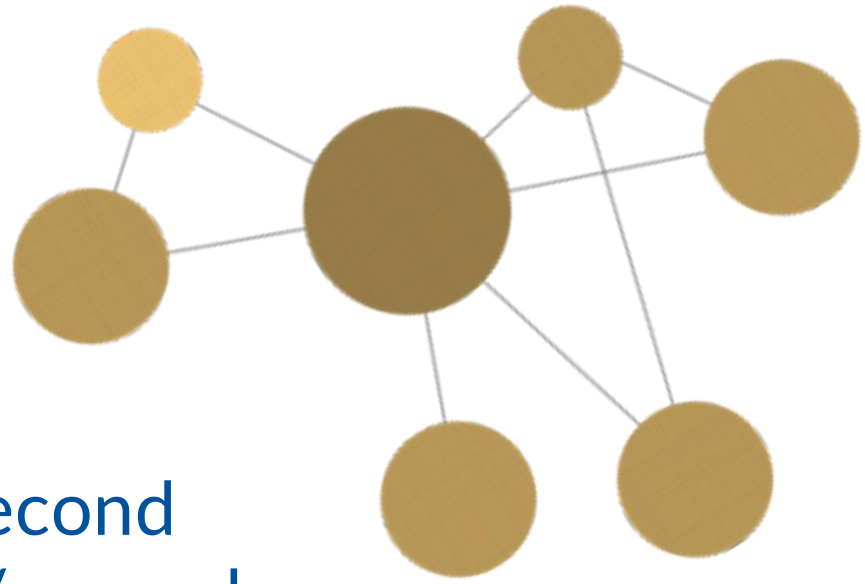
Latency

The Universal Truths

- The busier, something is, the higher its latency.
- When a connection is saturated, its latency skyrockets.
- The faster a link, the lower the processing latency.

Interconnects

What's faster? CAT6 or Fiber?



2 switches, 10 feet apart?

TECHNICALLY speaking…
Speed of light in a vacuum (c): 186,000 miles/second
Speed of light in earth air (.99c): 184,000 miles/second
Speed of light in fiber optic cable (.66c): 123,000 miles/second
Speed of transfer in twisted pair (.64c): 119,000 miles/second

Round trip latency: 30.8 nanoseconds for fiber, 31.8 for CAT6.

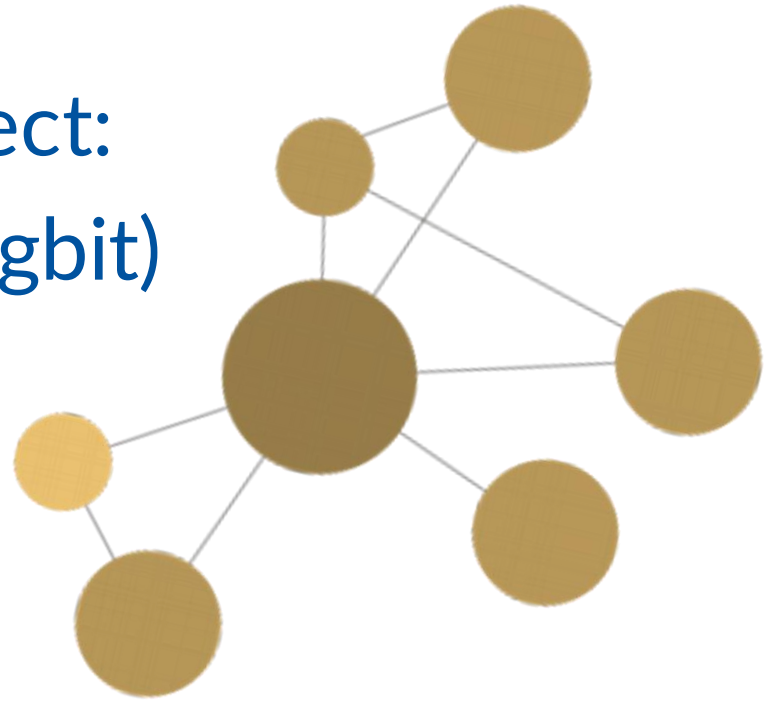
Transceivers add between 2 (rare) and 10 (common) microseconds of latency
Round trip latency: 40,030.8 nanoseconds for fiber, 31.8 for CAT6.
.04 milliseconds vs .0000318 milliseconds

Interconnects

None of that was realistic

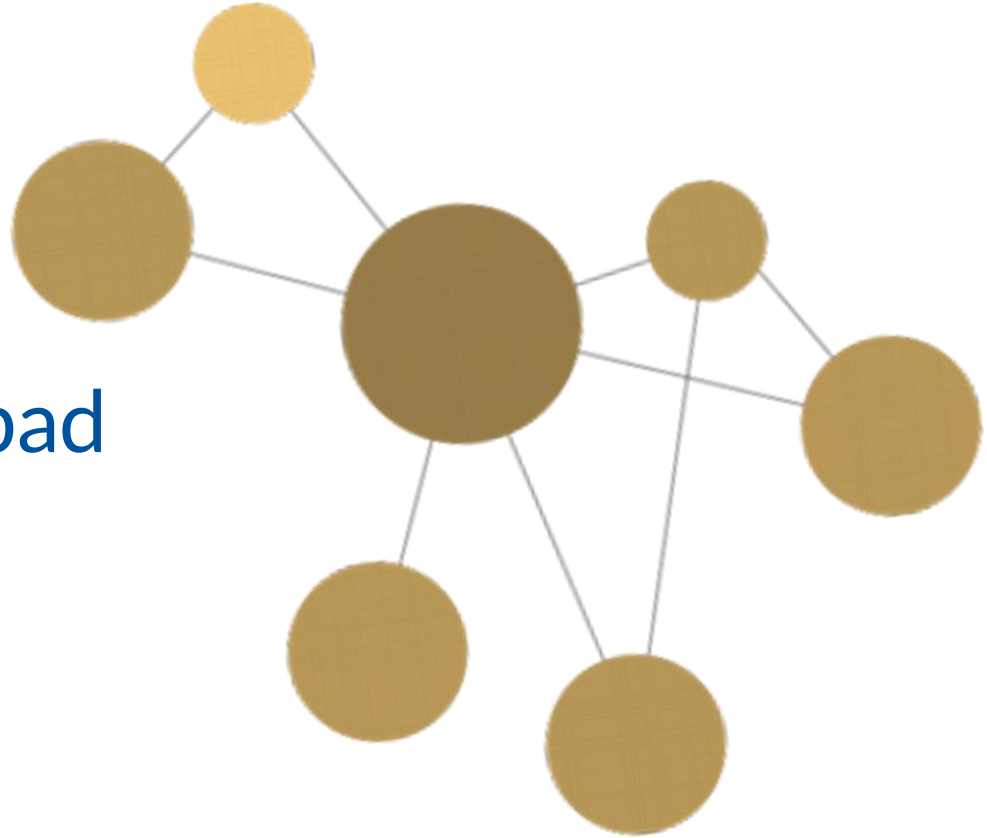
- If within distance -- without outside interference -- copper will give better latency, accounting only for physics.

- Many variables go into a simple interconnect:
- Connection speed (100mbit vs 1gbit vs 10gbit)
- Switching time
- NIC processing
- Compute processing



Interconnects

None of that was realistic



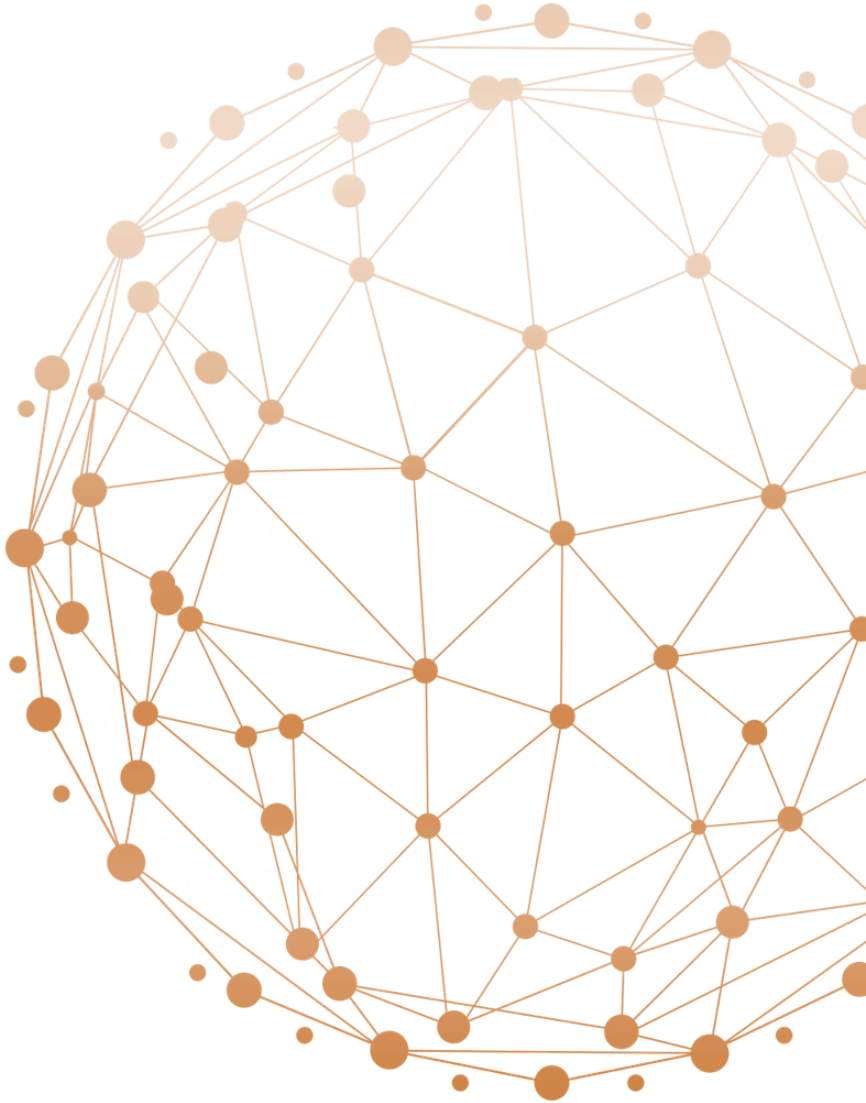
What about farther distances?

- Copper attenuates – repeaters are bad
- Cross talk
- Leased lines have equipment in line
- Cables don't run in straight lines
- Satellites
  - Nearly the speed of light
  - Adds a HUGE distance 12,000-24,000 miles per trip

Generic Network Woes

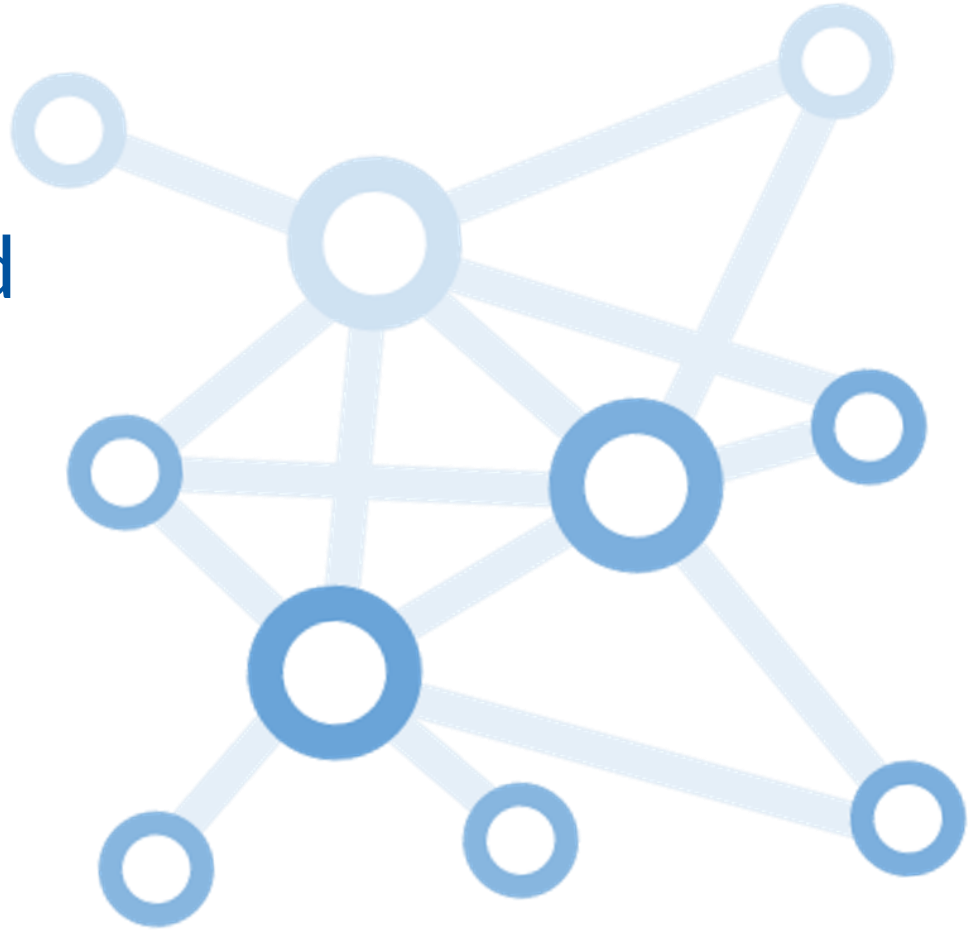
TCP/IP, Ethernet and Duplex and speed

- TCP/IP, Ethernet
  - Encapsulation/decapsulation
  - NIC offload capabilities can help
  - Jumbo Frames CAN help, but not necessarily
  - FCoE less overhead than iSCSI
  - FC less overhead than FCoE
- Duplex and Speed still a concern
  - Especially prevalent on leased lines
- Spanning tree convergence



Generic Network Woes

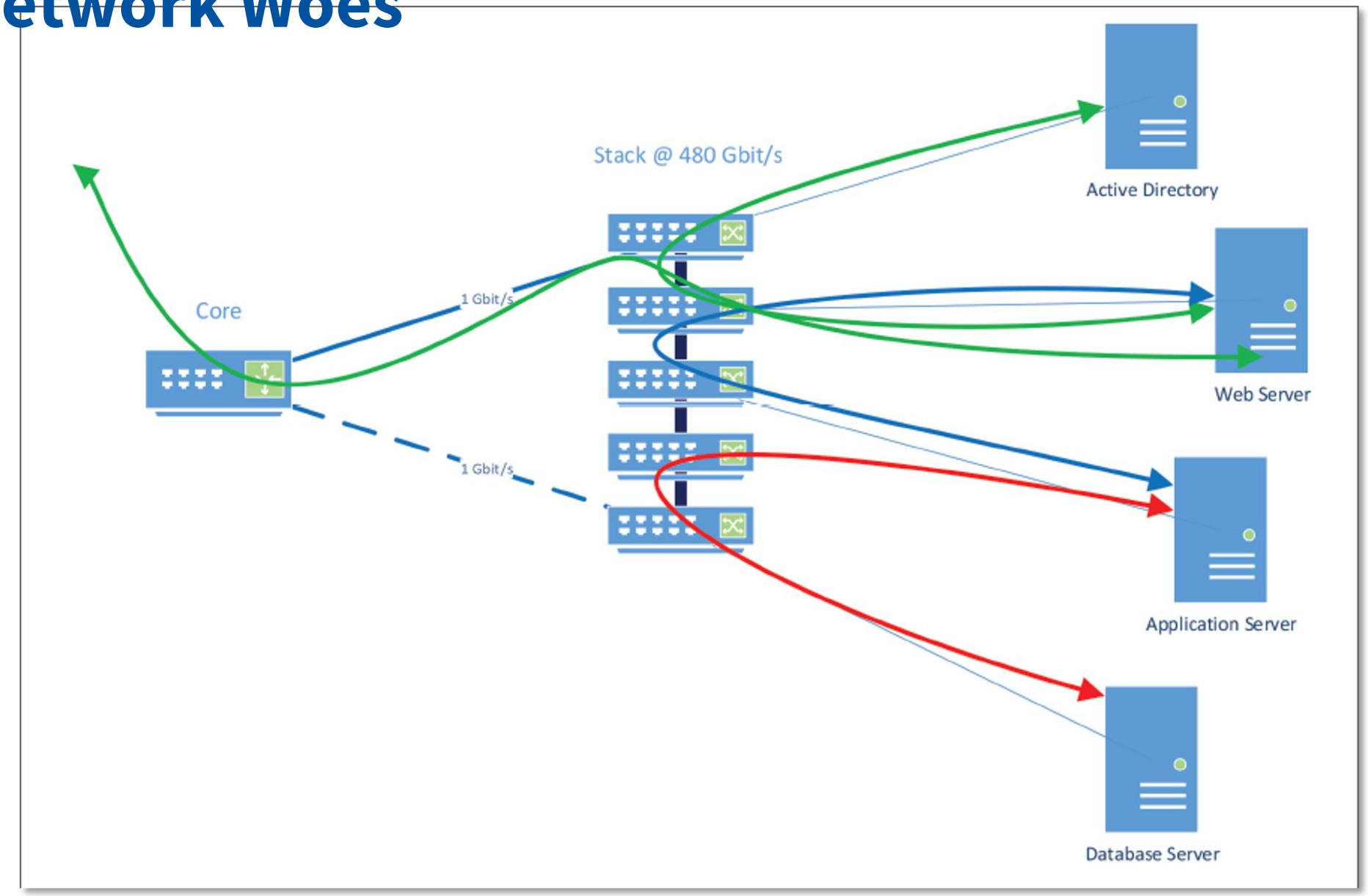
Routing



- Modern networks heavily subnetted
  - Where is the routing happening?
  - Switches
  - Routers
  - How many hops away?
  - How overcommitted are those uplinks?
  - Full links cause latency
- Minor changes can cause CPU-based routing

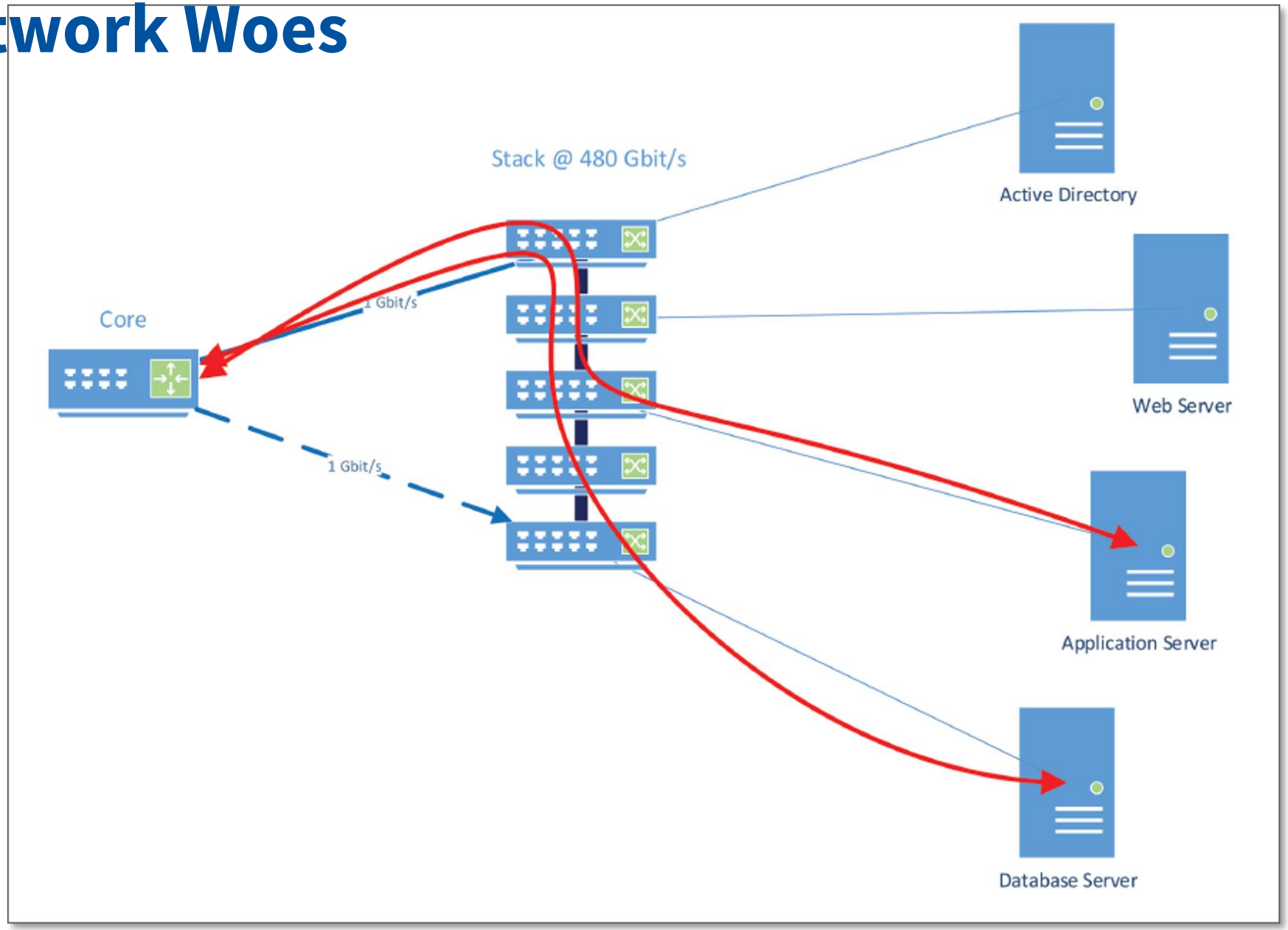
Generic Network Woes

Routing



Generic Network Woes

Routing



Generic Network Woes

The Internet

- Uncontrolled chaos
- Thousands of paths
  - none care about you
- Drops all QoS
- Paths can regularly flip
- Dozens of providers in every path
- Makes some on premise things feel slow
- Firewalls can add a lot of latency

Hardware

Motherboard CPU & RAM

- Not all PCI-E links are created equal
  - Generation 1, 2, 3, 4
  - Electrical wiring vs physical slots (x16 slots aren't necessarily)
  - Slot overcommitment – Storage controller with 16 SSDs in a single x8 slot
- NUMA – QPI half as fast as direct memory
- Properly lay out your DIMMs (RTFM)

Operating Systems

Hypervisor

- CPU overcommitment/CPU ready
- vNUMA
- Memory compression/swapping
- Virtualization drivers
- vSwitch uplink overcommitment
- Localized caching to the host
- Routing between VMs [NSX?]

Operating Systems

Physical OS

- Keep firmware and BIOS up to date
- Keep drivers up to date
- C states?
- Turboboost?
- Hyperthreading?

Storage

Generic

- Vendors ~~lie~~, ~~modify~~ ~~the~~ ~~truth~~, obfuscate context to their benefit.
- Context is insanely important
- Large focus on storage latency now
- Historically storage
  - Has had highest latency
  - Has had fewest latency advances
  - Has been least designed for performance

Storage

Generic

- Watch for:
  - IOPs without block size
  - IOPs without avg/max latency
- Spinning disks
  - 7.2k high latency
  - 10k & 15k much lower, 15k replaced by SSD
  - Short stroking – finally almost dead

Storage

SSD

- Amazing at reading
- Good at writing … usually
- Write amplification
- Pages and blocks
- Writes wear out SSDs
- Important to have disks and arrays designed for writes with SSDs

Storage

RAID

- RAID 1/10
  - Half the write performance always
- RAID 5/50
  - Between N-1 and 25% write performance
  - N-1 Read
- RAID 6/60
  - Between N-2 and 16% write performance
  - N-2 Read
- Rebuilds!

Storage

Features

- Dedupe and Compression add latency
- Auto tiering can sometimes compensate for latency
  - Speed of tiering important
  - Process of tiering can slow down performance
- Caching
  - Good way to boost slower storage
  - Is the cache ever lost?
  - Penalty for rebuilding it

User Experience

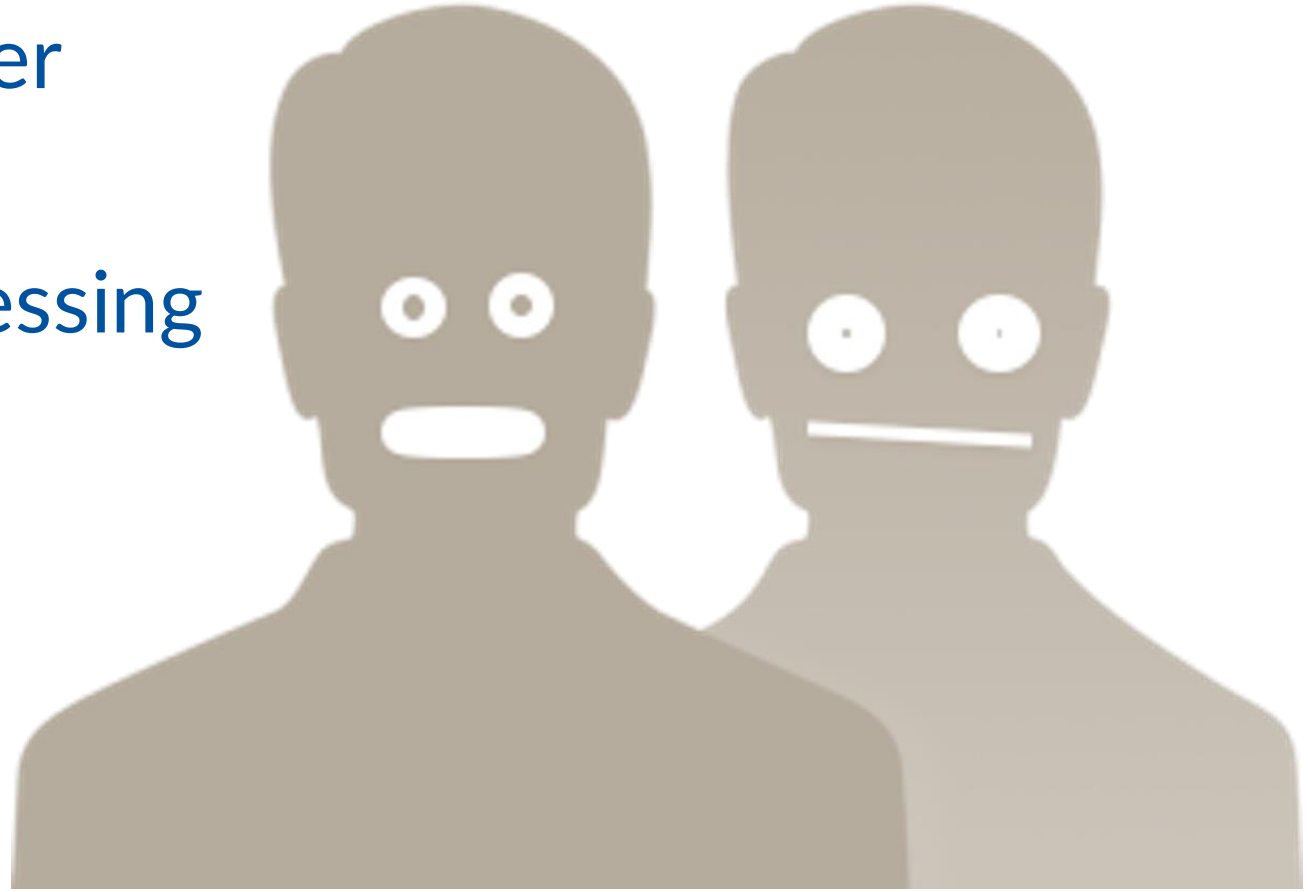
General

- Aggregate of all latency everywhere
- 100ms total latency where it's noticed in direct interaction
- Includes everything from user endpoint through datacenterDifferent concerns for different types of users
- Application type
- Chatty client/server
- Server based

User Experience

Generically

- Desktop Speed
- Network speed to datacenter
- Datacenter network
- All servers involved in processing
- Storage
- Path back



User Experience

Bottleneck based on location

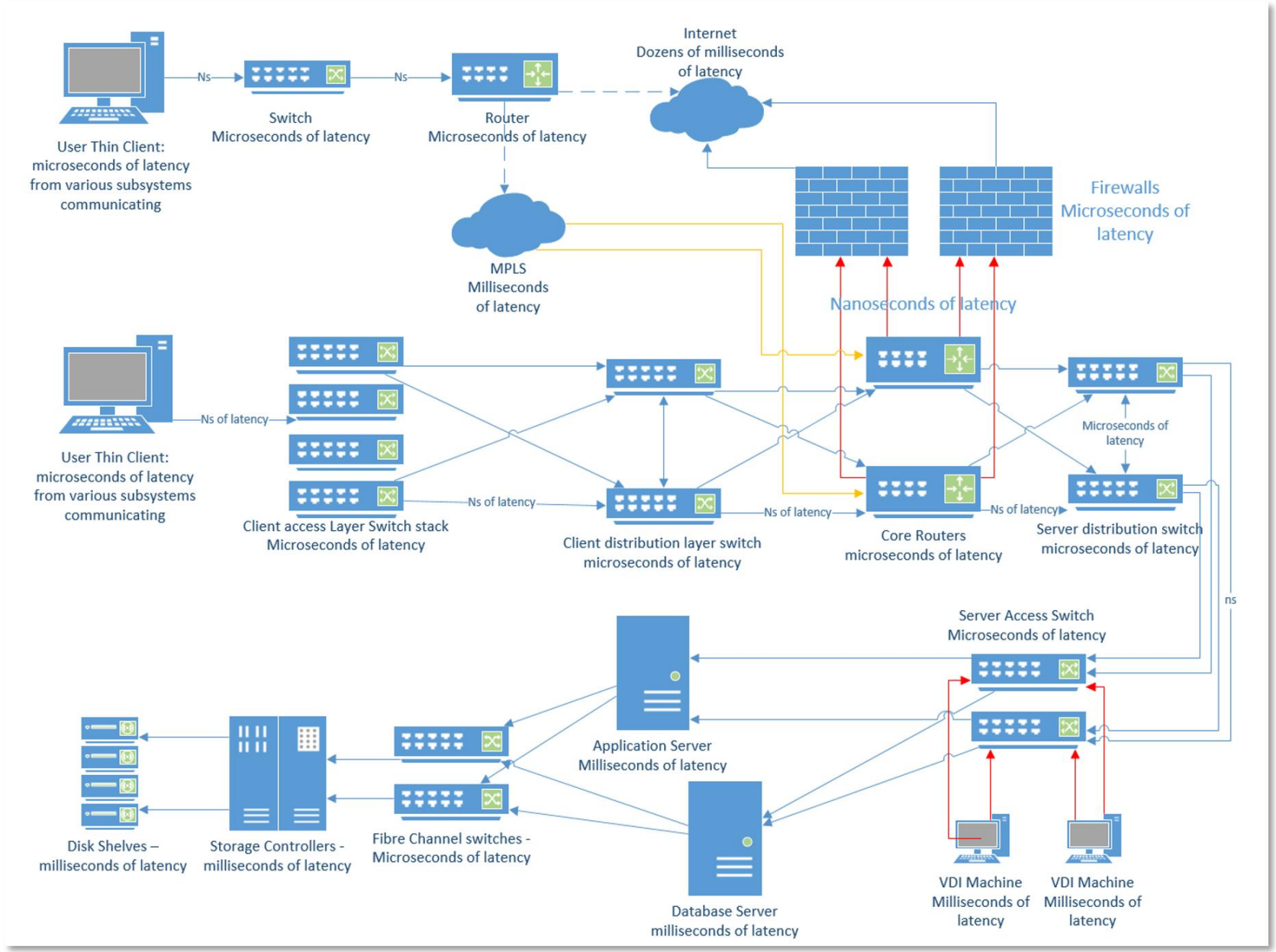
- On premise
  - Application processing
  - Storage
  - Fixed by better apps and flash often
- MPLS/VPN
  - Path to datacenter
  - Chatty client/server applications
  - Application processing
  - Storage
  - Fixed by VDI, better apps often

User Experience

VDI / Remote Sessions

- Helps many remote issues
- Remedies chatty apps
- Can mitigate impact of high latency link
- 100 ms still gold standard for whole user experience
- Pay attention to jitter
- Can build credit towards other latencies

User Experience



The Cloud

Panacea … or latency nightmare?

- No benefit for in office or on the road
- Internet latency important
- Firewall latency important
- Geolocation?
- Cloud's latency important (full stack in cloud)
- Hybrid (resources close to each other that talk a lot)
- Oversaturated pipes

Thank You / Questions?

Brent.Earls@mirazon.com

Mirazon.com