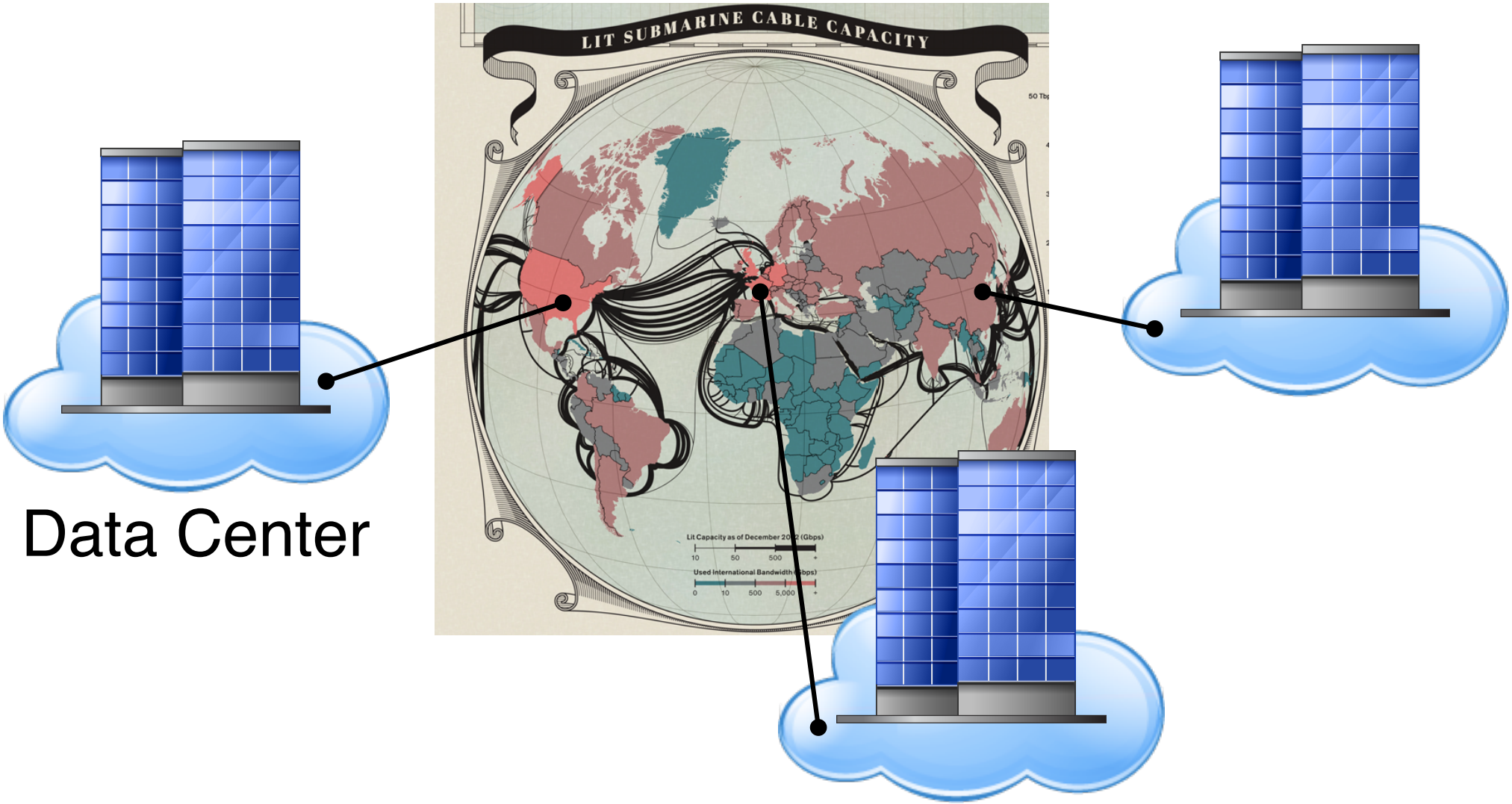
Distributed Systems CS6421

Storage: Scale and Fault Tolerance

Prof. Tim Wood

Our picture of the cloud…

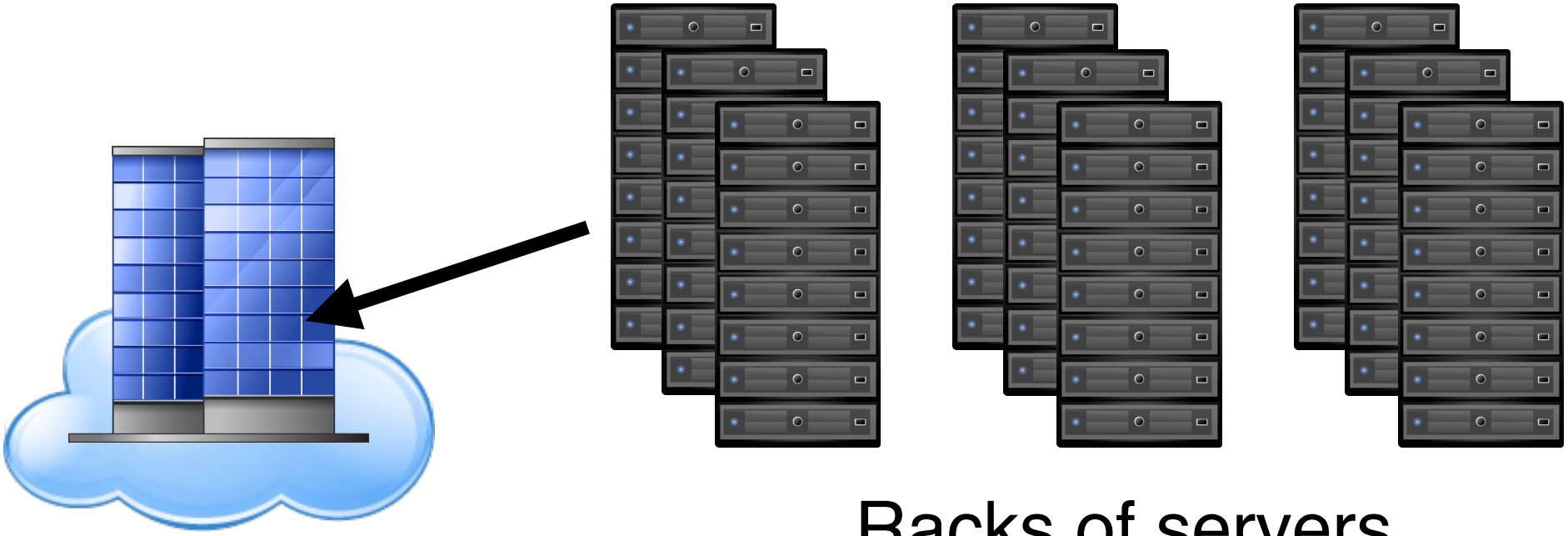
Data centers connected around the world…



Data Center

Our picture of the cloud…

Filled with lots of servers…

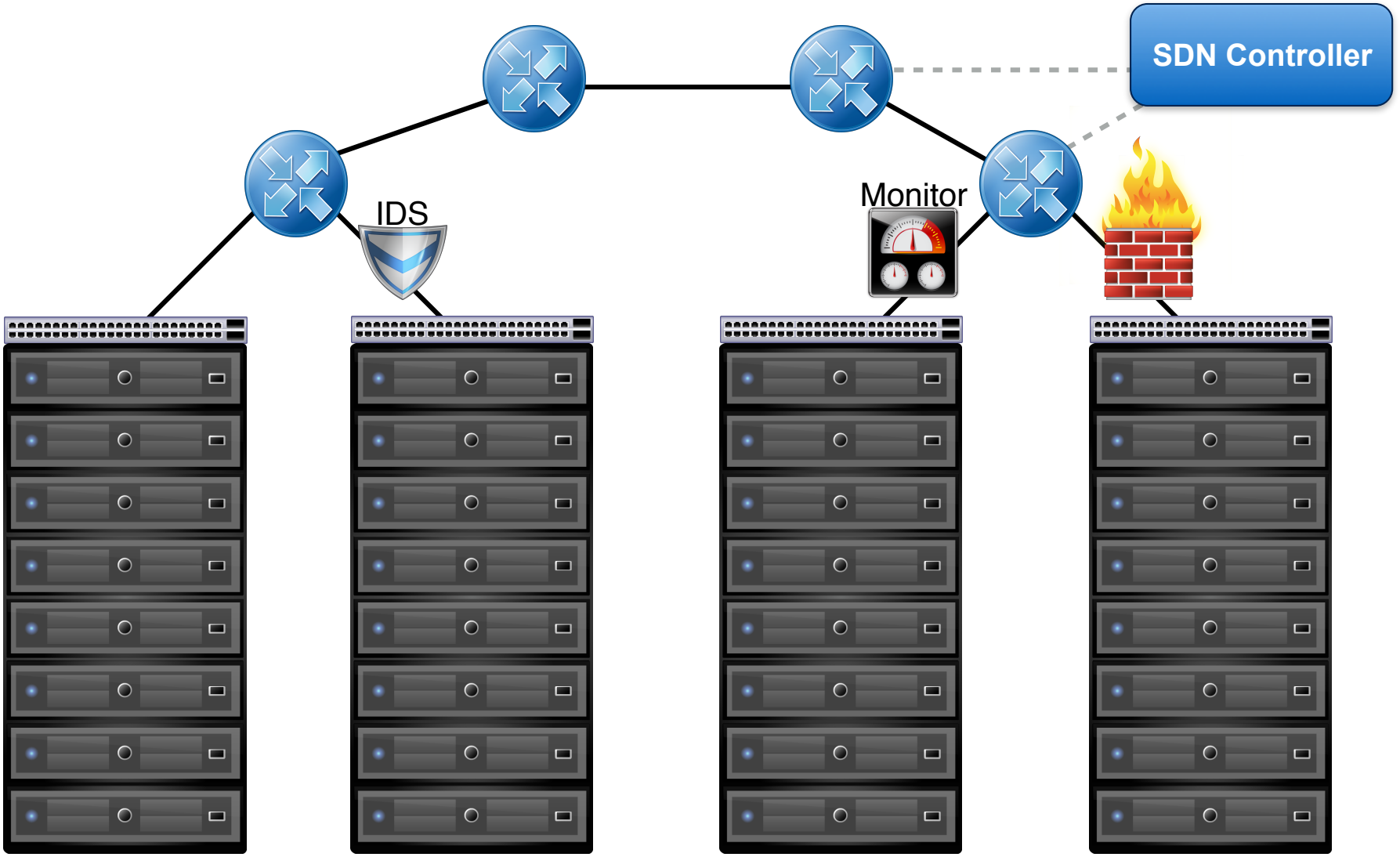


Data Center

Racks of servers

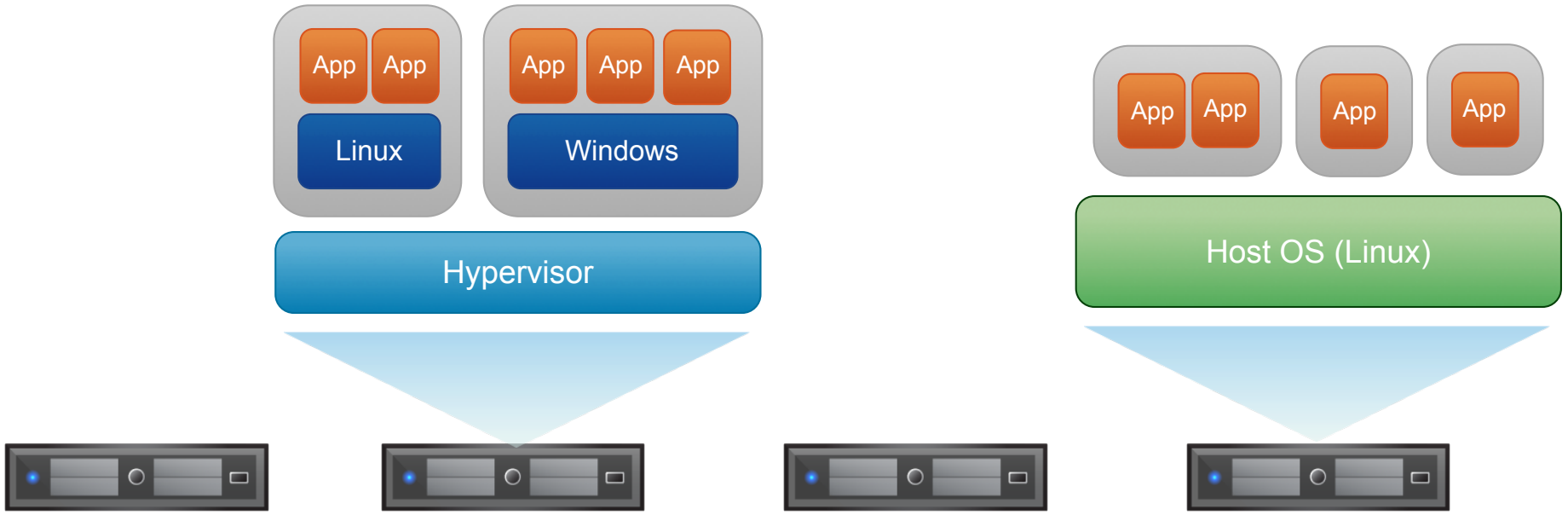
Our picture of the cloud…

Connected with routers, switches, and middle boxes



Our picture of the cloud…

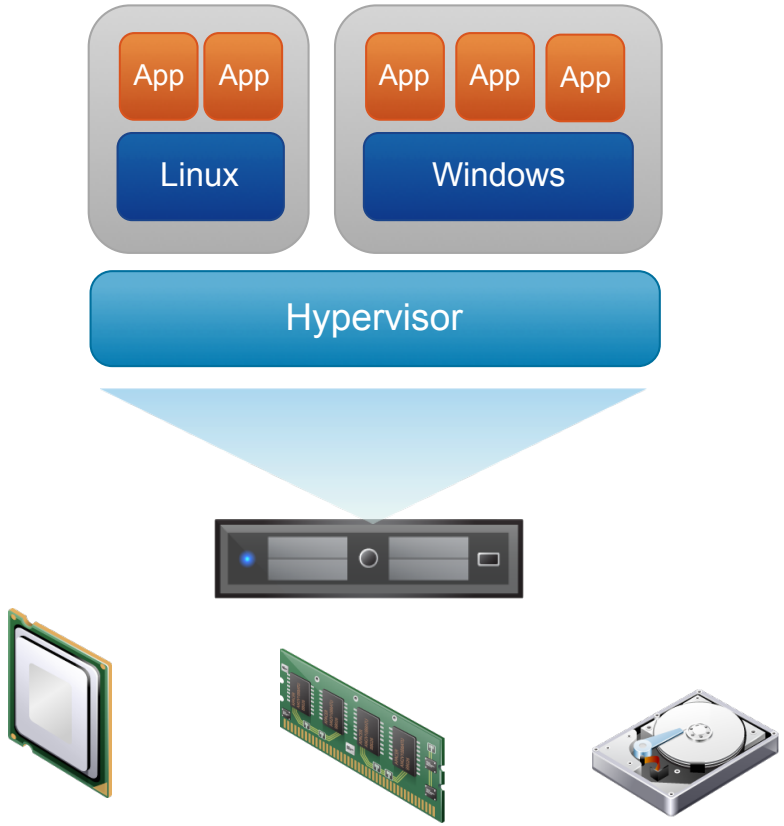
Each hosting VMs or containers



Resources?

CPU, RAM, and Disks

- How is storage different?



Numbers you should know…

How long to…

```
L1 cache reference ..................... 0.5 ns
Branch mispredict ....................... 5 ns
L2 cache reference ...................... 7 ns
Mutex lock/unlock ..................... 25 ns
Main memory reference ................. 100 ns
Compress 1K bytes with Zippy ............. 3,000 ns  =   3 µs
Send 2K bytes over 1 Gbps network ....... 20,000 ns  =  20 µs
SSD random read ........................ 150,000 ns  = 150 µs
Read 1 MB sequentially from memory ..... 250,000 ns  = 250 µs
Round trip within same datacenter ...... 500,000 ns  = 0.5 ms
Read 1 MB sequentially from SSD* ..... 1,000,000 ns  =   1 ms
Disk seek ........................... 10,000,000 ns  =  10 ms
Read 1 MB sequentially from disk .... 20,000,000 ns  =  20 ms
Send packet CA->Netherlands->CA .... 150,000,000 ns  = 150 ms
```

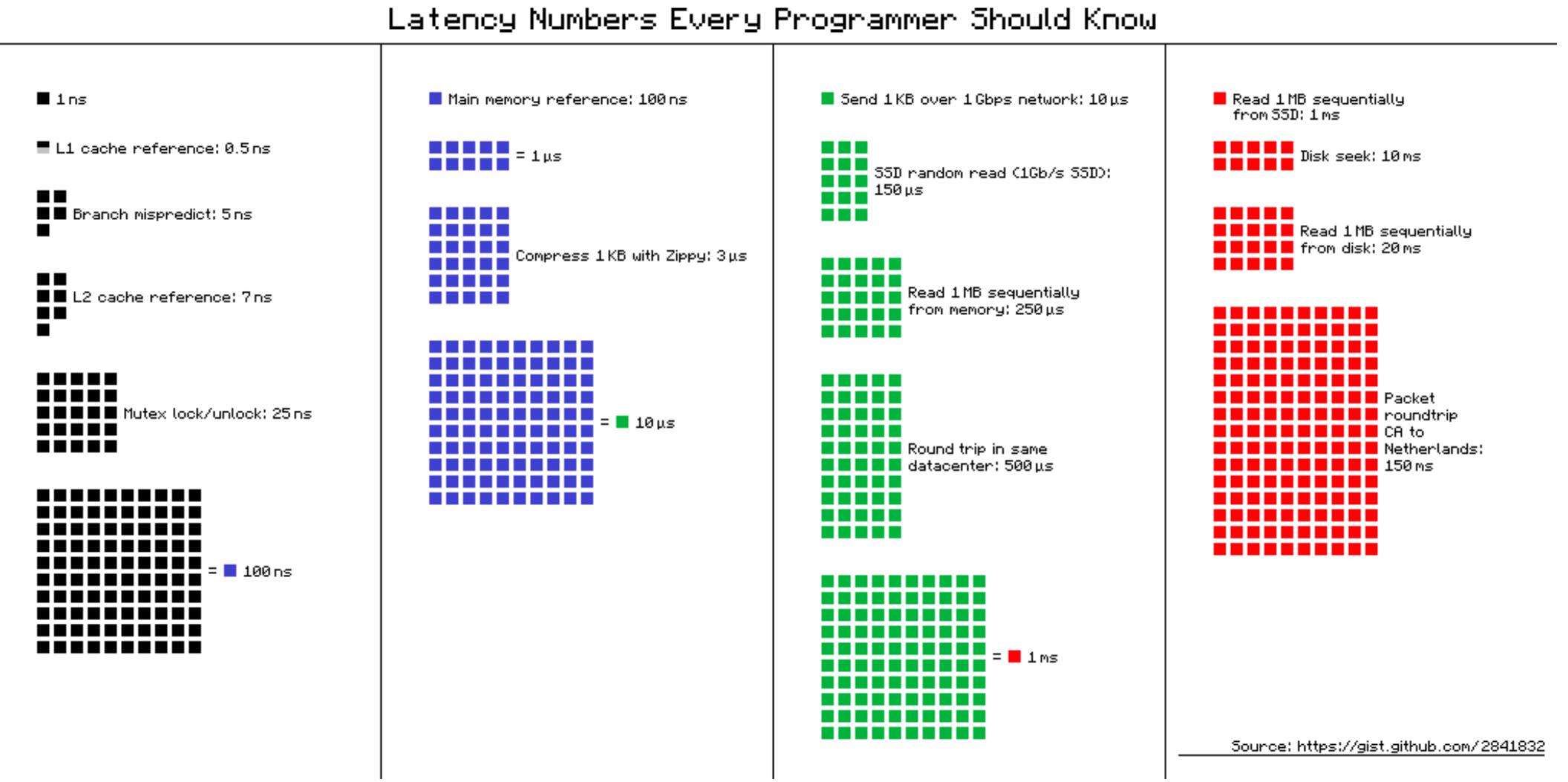
Numbers you should know…

How long to…

```
L1 cache reference ..................... 0.5 ns
Branch mispredict ....................... 5 ns
L2 cache reference ...................... 7 ns
Mutex lock/unlock ..................... 25 ns
Main memory reference ................... 100 ns
Compress 1K bytes with Zippy ............. 3,000 ns  =   3 µs
Send 2K bytes over 1 Gbps network ....... 20,000 ns  =  20 µs
SSD random read ........................ 150,000 ns  = 150 µs
Read 1 MB sequentially from memory ..... 250,000 ns  = 250 µs
Round trip within same datacenter ...... 500,000 ns  = 0.5 ms
Read 1 MB sequentially from SSD* ..... 1,000,000 ns  =   1 ms
Disk seek ........................... 10,000,000 ns  =  10 ms
Read 1 MB sequentially from disk .... 20,000,000 ns  =  20 ms
Send packet CA->Netherlands->CA .... 150,000,000 ns  = 150 ms
```

https://gist.github.com/hellerbarde/2843375

Numbers you should know…



also: https://people.eecs.berkeley.edu/~rcs/research/interactive_latency.html

Scale it up…

After multiplying everything by a billion…

L1 cache reference	0.5 s	One heart beat (0.5 s)
L2 cache reference	7 s	Long yawn
Main memory reference	100 s	Brushing your teeth
Send 2K bytes over 1 Gbps network	5.5 hr	From lunch to end of work day
SSD random read	1.7 days	A normal weekend
Read 1 MB sequentially from memory	2.9 days	A long weekend
Round trip within same datacenter	5.8 days	A medium vacation
Read 1 MB sequentially from SSD	11.6 days	Waiting 2 weeks for a delivery
Disk seek	16.5 weeks	A semester in university
Read 1 MB sequentially from disk	7.8 months	Almost producing a new human being
The above 2 together	1 year	
Send packet CA->Netherlands->CA	4.8 years	Going to university for BS degree

???

Scale it up…

After multiplying everything by a billion…

L1 cache reference	0.5 s	One heart beat (0.5 s)
L2 cache reference	7 s	Long yawn
Main memory reference	100 s	Brushing your teeth
Compress 1K bytes with Zippy	50 min	One episode of a TV show
Send 2K bytes over 1 Gbps network	5.5 hr	From lunch to end of work day
SSD random read	1.7 days	A normal weekend
Read 1 MB sequentially from memory	2.9 days	A long weekend
Round trip within same datacenter	5.8 days	A medium vacation
Read 1 MB sequentially from SSD	11.6 days	Waiting 2 weeks for a delivery
Disk seek	16.5 weeks	A semester in university
Read 1 MB sequentially from disk	7.8 months	Almost producing a new human being
The above 2 together	1 year	
Send packet CA->Netherlands->CA	4.8 years	Going to university for BS degree

Accessing 1MB data

Data in RAM:

- 3 days (250 **micro**seconds)
- 60 GBps

Data in SSD:

- 2 weeks (1 **milli**second)
- 1 GBps

Data in HDD:

- a year (20 **milli**seconds)
- 100 MBps

Send 1MB over 10Gbps network:

- length of this class (2 **milli**seconds)
- 1.25 GBps

Networked Storage

The added cost of using the network is relatively low

What are the benefits of using remote storage instead of local?

Storage Services

Block storage (EBS)

- Access to the raw bytes of a remote disk
- Unit of access: disk block (4KB)
- Mount as the disk for a VM
- Pros/Cons?
  - Different types of underlying storage (SSD vs HDD)
  - Pay per GB but I need to reserve the space in advance, number of IOPs?
  - Limited to 16TB per disk

Storage Services

Block storage (EBS)

Object storage (S3, DynamoDB)

- Get and Put “objects” in remote storage
- Unit of access: a full object
- Store static web content, data sets
- Pros/Cons?
  - Pay per object based on size, also per request, net BW?
  - Limit 5 TB per object

Storage Services

Block storage (EBS)

Object storage (S3, DynamoDB)

Database (RDS)

- Store structured rows and columns of data
- Unit of access: SQL query
- Web applications requiring transaction support
- Pros/Cons?
  - AWS handles management

Implementation?

How would you build a Block/Object store?

- What abstraction layers?
- What should the interface look like?
- What traits do you optimize for?

What price could you sell it for?