Building warehouse-scale computers

or … what's it like to supply *exponential growth*

john wilkes 2018-10

Google Cloud Platform

Video: https://youtu.be/m7uIG8qFGMI

~~You're all~~

Some of you are

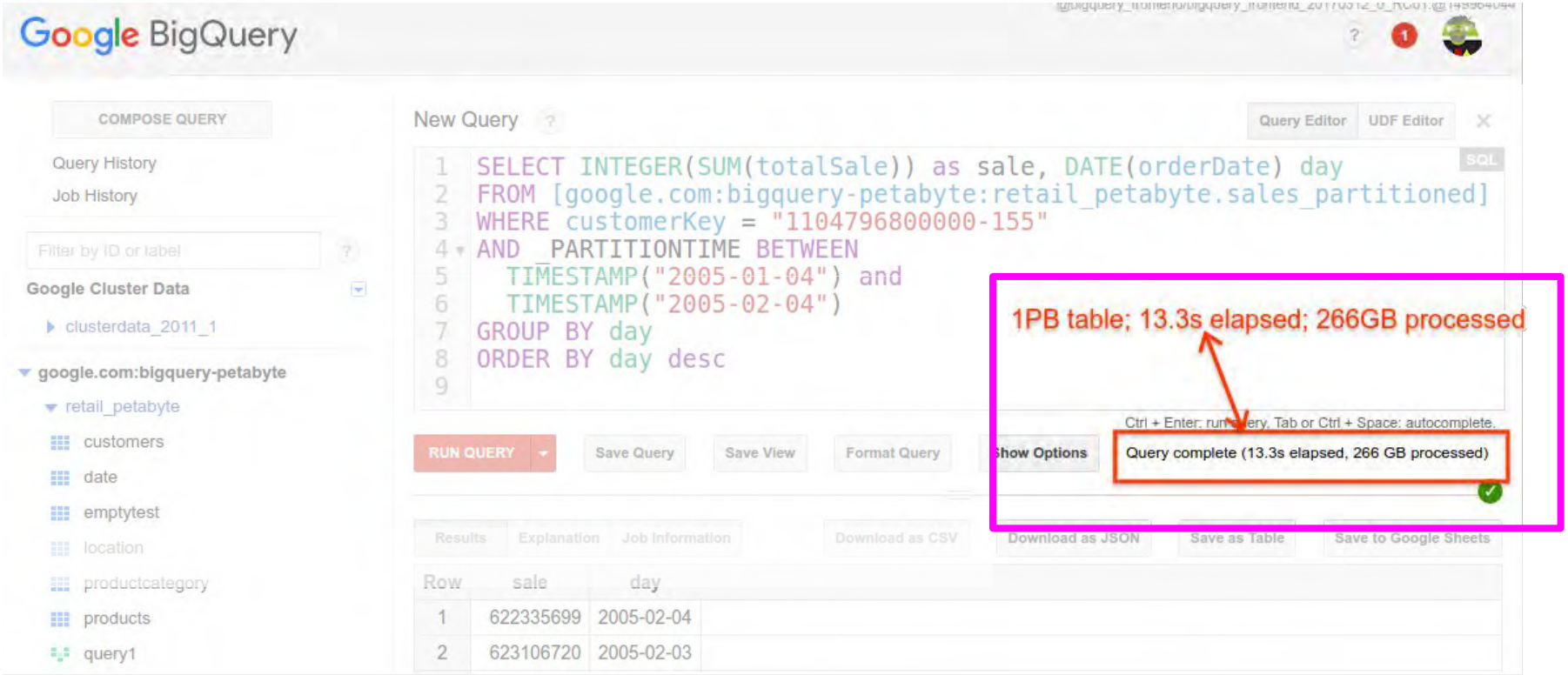
thinking

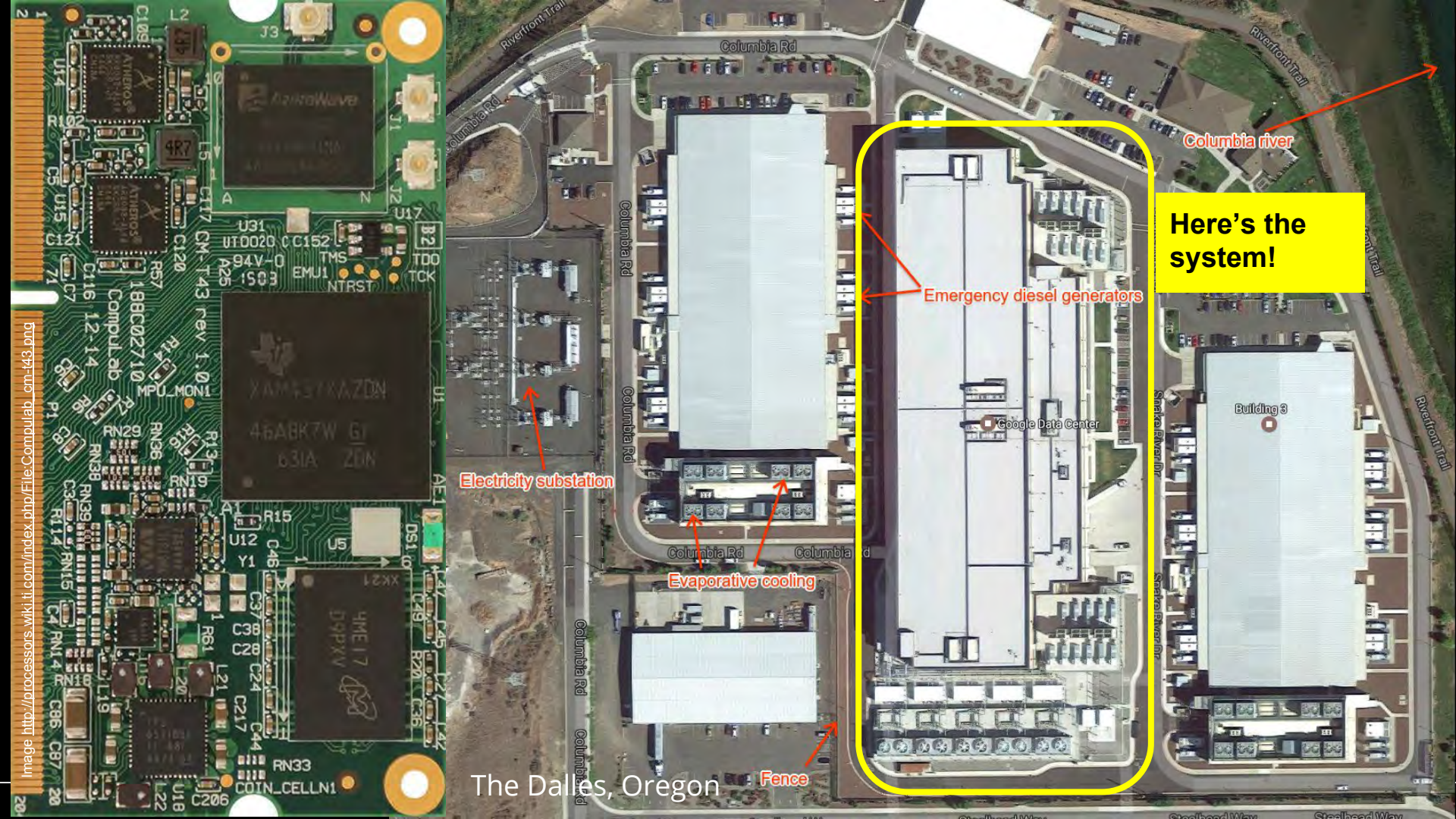
too small

Scale has been the single most important force driving changes in system software over the last decade

– Technical perspective: Is scale your enemy, or is scale your friend?
John Ousterhout, CACM 54(7):110, July 2011.

The kind of things we like to be able to do




Here's the system!
Columbia river
Emergency diesel generators
Electricity substation
Evaporative cooling
Fence
Google Data Center
Building 3
Columbia Rd
Riverfront Trail
The Dalles, Oregon
Image http://processors.wiki.ti.com/index.php/File:Compulab_cm-t43.png

What if …
the “**system**” included not just the computers,
but the network fabric and the WAN endpoints

Council Bluffs, Iowa

and the cooling system

and the building-management system

St Ghislain, Belgium

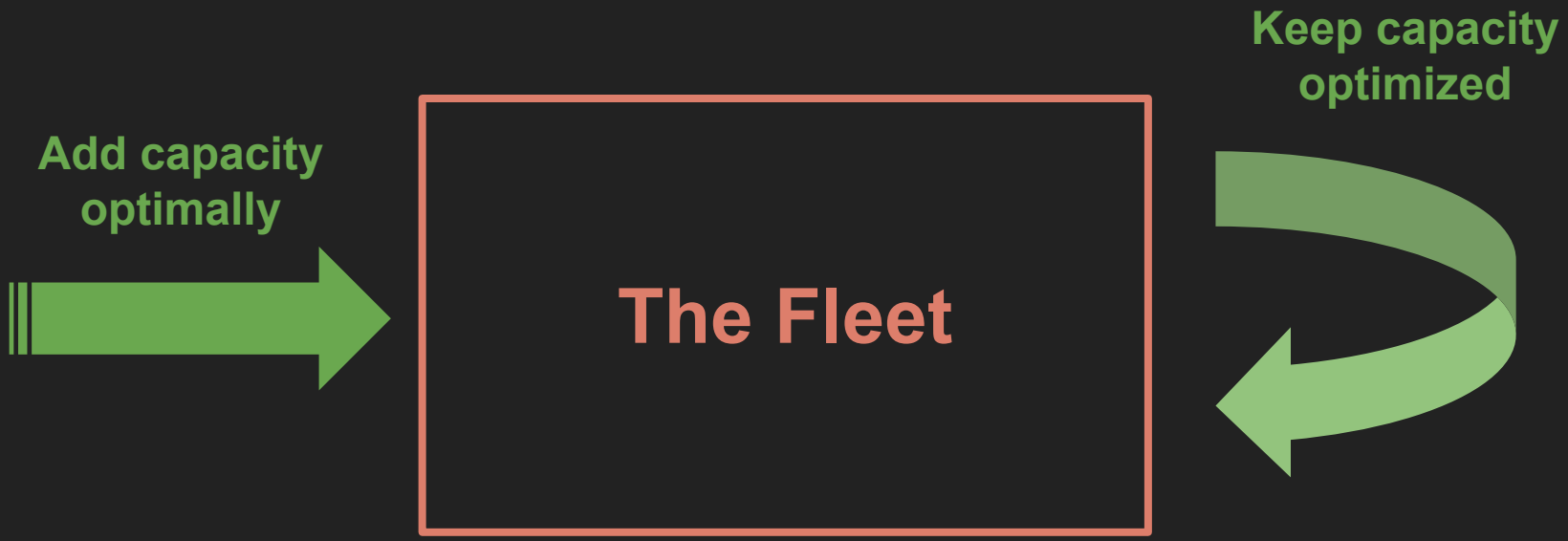
and the power system
and the technicians

Backup generator at
St. Ghislain, Belgium

In fact … the entire **warehouse scale computer**

By the way … it costs O($200M). No pressure.

Key challenges



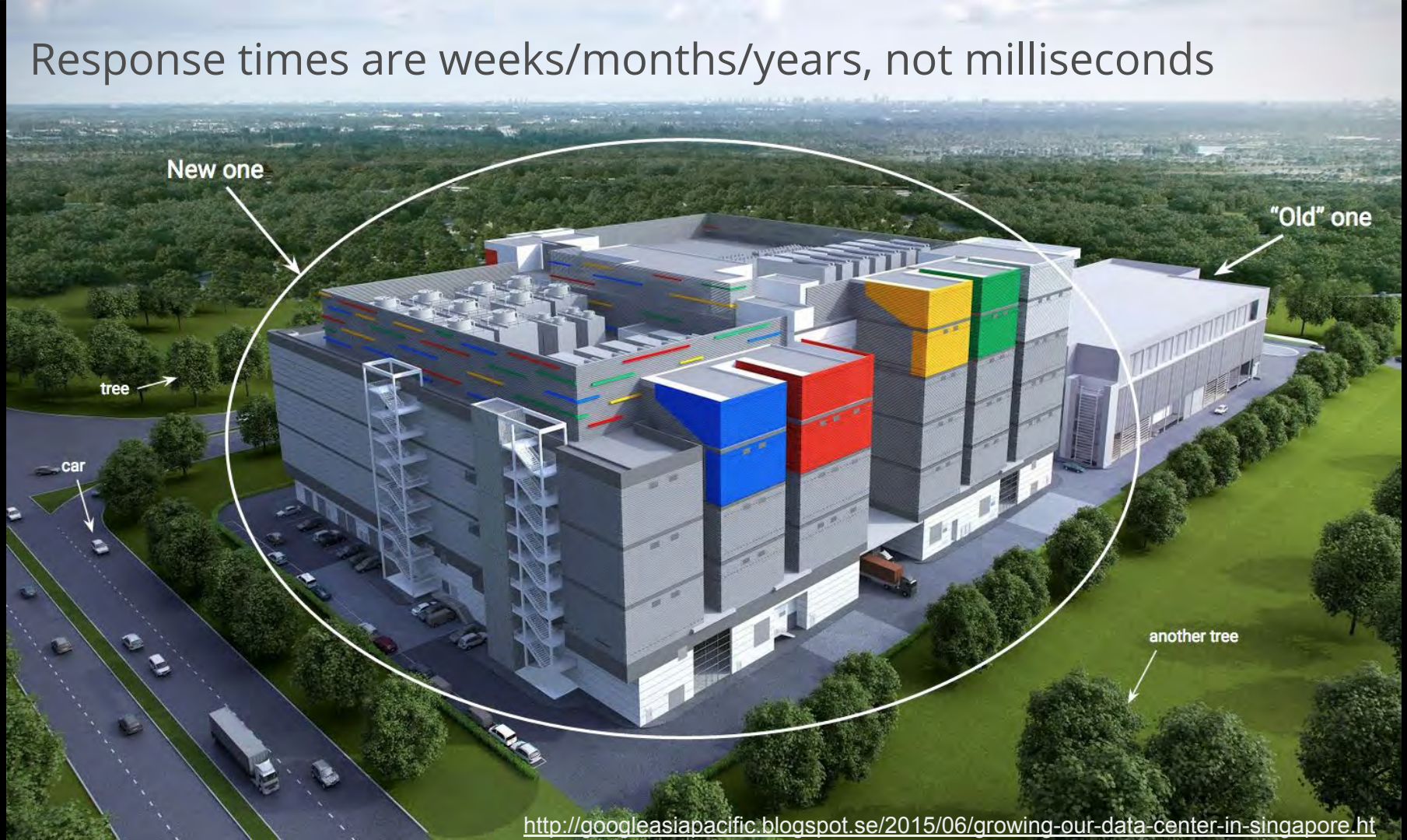
Planning for compute resources

All have multiple generations/options

- CPU
- RAM
- disk, SSD – capacity, performance

- new NVRAM thingies …
- accelerators (complication: a wide variety)
- inter- and intra-datacenter networking
- power
- datacenters
- land, water, sewage
- …

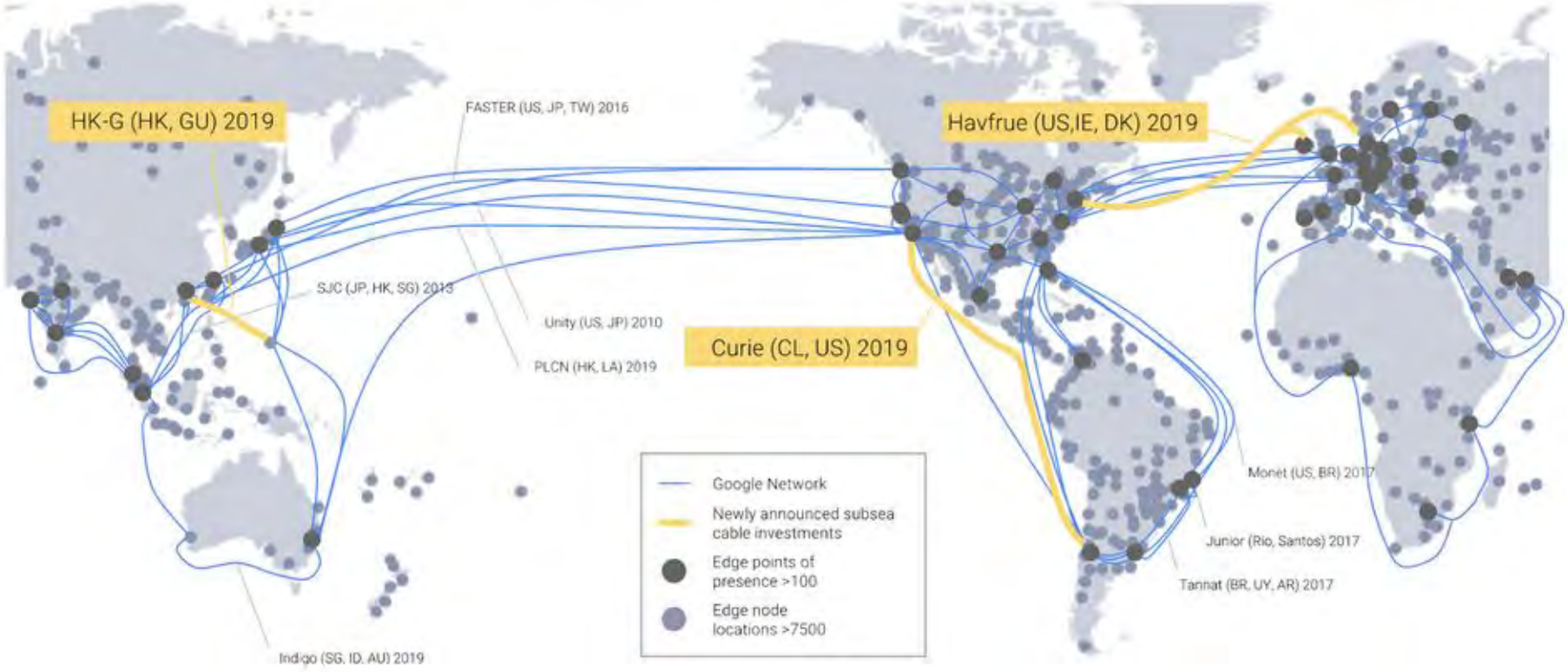
Response times are weeks/months/years, not milliseconds



http://googleasiapacific.blogspot.se/2015/06/growing-our-data-center-in-singapore.ht

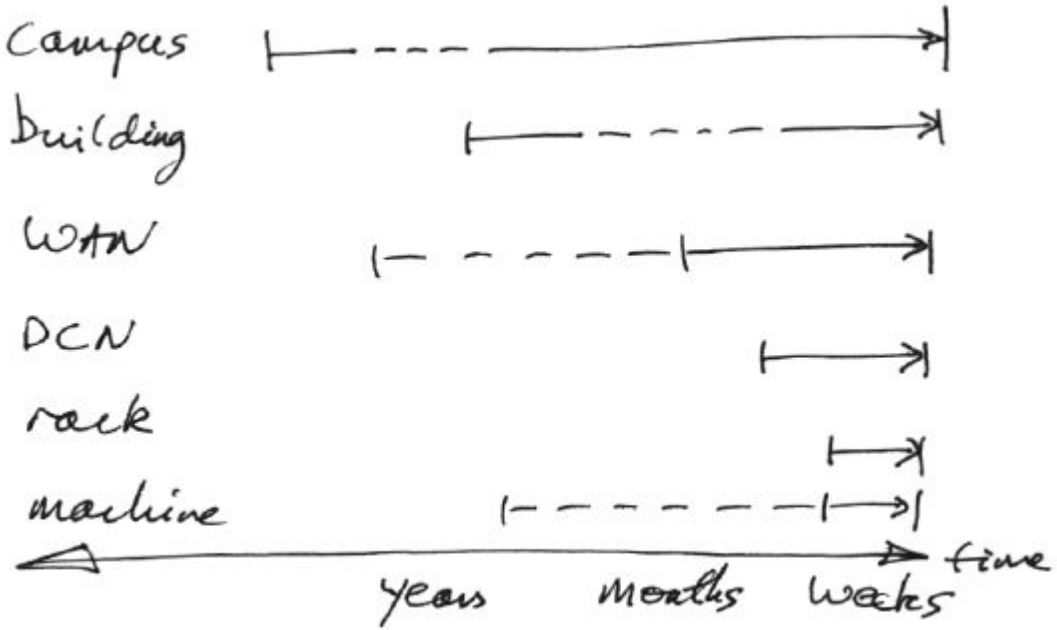
Google Network

The largest cloud network, comprised of more than 100 points of presence



How long does it take to build capacity?

Note: variance is as bad as delay



Planning for compute resources

What kind of machine to buy?

- different groups want different things (e.g., search, Cloud)
- ⇒ MotD + customizations

Idea: reuse machines when they get handed down

- when is it worthwhile? (price of power? room for expansion?)

Also: do we have … space? power? networking? budget?

- models; what-if analyses; uncertainty (demand, supply)
- objective function: *total cost-of-ownership*

Planning for compute resources: Total Cost of Ownership (TCO)

what do you mean by “**total**”?

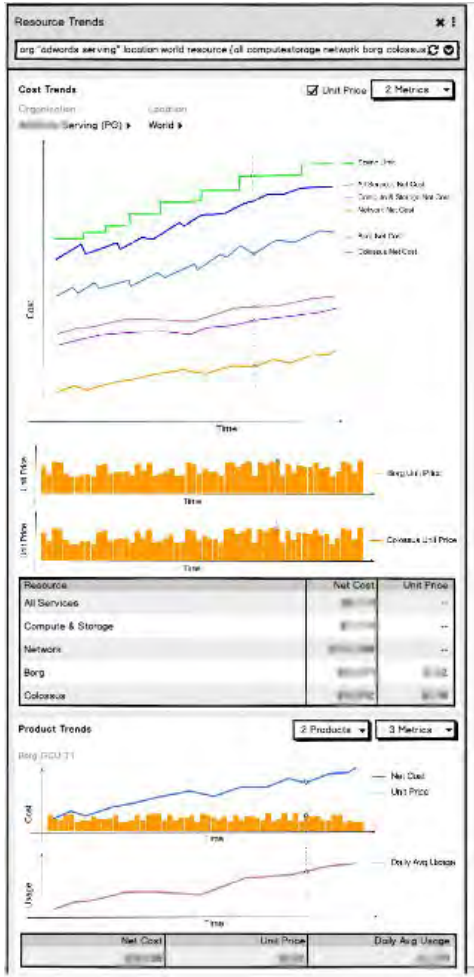
- average? over what? continent? time?
- depreciation schedule?

what do you mean by “**cost**”?

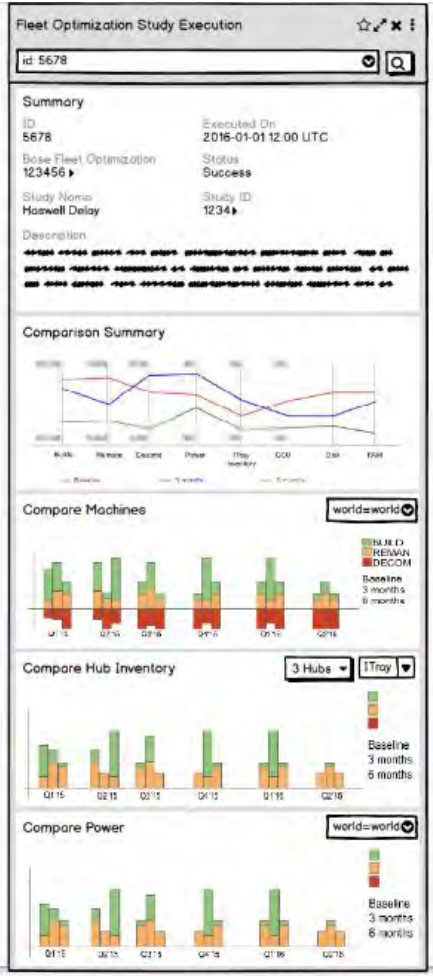
- initial purchase price, or the average over time?
- is delivery or installation included? what if they are being reused?

what do you mean by “**ownership**”?

- machines can be transferred/given/sold/break
- who “owns” a machine running a shared service for a customer?

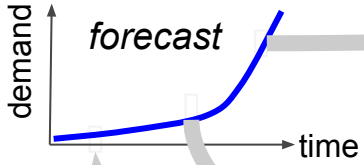

Resource Trends
Cost Trends
Unit Price
2 Metrics
Organization
Location
Serving (PG)
World
Time
Cost
Unit Price
Resource
Net Cost
Unit Price
All Services
Compute & Storage
Network
Borg
Colossus
Product Trends
2 Products
3 Metrics
Net Cost
Unit Price
Daily Avg Usage
Usage
Time


Inventory
group:it-flex-pool location:region=asia service:bigtabl
Summary
Group
it-flex-pool
Daily Cost
Health
Healthy
Location
Region Asia
Footprints
Scaling
1 needs attention
2 executing
10 recent
Service
Bigtable
Registration Status
Open
Product Metrics
4 Metrics
Product
Capacity
Commitment
Available
Available %
Servers
554
499
55
10%
GCU
28%
RAM
61%
Low Latency
10%
Throughput
28%
Disk
61%
SSD
10%
Capacity Breakdown
Disk
Allocated
Metric Timelines
Total Cost


Fleet Optimization Study Execution
id: 5678
Summary
ID
5678
Executed On
2016-01-01 12:00 UTC
Base Fleet Optimization
123456
Status
Success
Study Name
Haswell Delay
Study ID
1234
Description
Comparison Summary
Compare Machines
world=world
BUILD
REMAIN
DECOM
Baseline
3 months
6 months
Compare Hub Inventory
3 Hubs
1 Troy
Baseline
3 months
6 months
Compare Power
world=world
Baseline
3 months
6 months

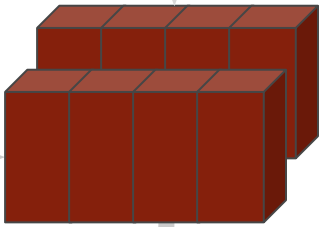

Order
id:123454321
Summary
ID
12345432
Supplier
machine-service-so
Service
PA-Specified Machine
Consumer
ads-pantry
Type
Create
Purpose
Product launch
Location
Cluster io
Category
None
Requested start date
2016-12-31
Relationship
12345 Growth for my PA
Description
lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex
Status
Draft
Submitted
Validated
Policy Evaluation
Passed budget check, Passed capacity check, Passed every other check check check..
Products
Requested
Ordered
Planned
Delivered
GTIM00003
3k
3k
GTIM00004
4k
4k
GTIM00005
4k
4k
Open Forecast for Lego Inventory
Cost
Monthly Breakdown
Ordered
Planned
Delivered
Pending Delivery
Dec 2016
Jan 2017
Feb 2017
Mar 2017
Apr 2017

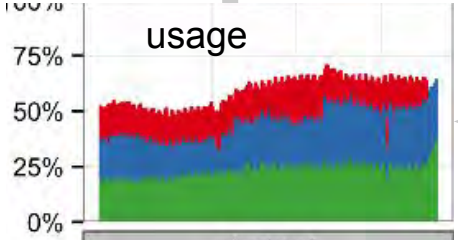
A simplified overview



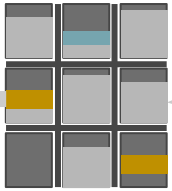
sites, data centers, power

network, machines, storage





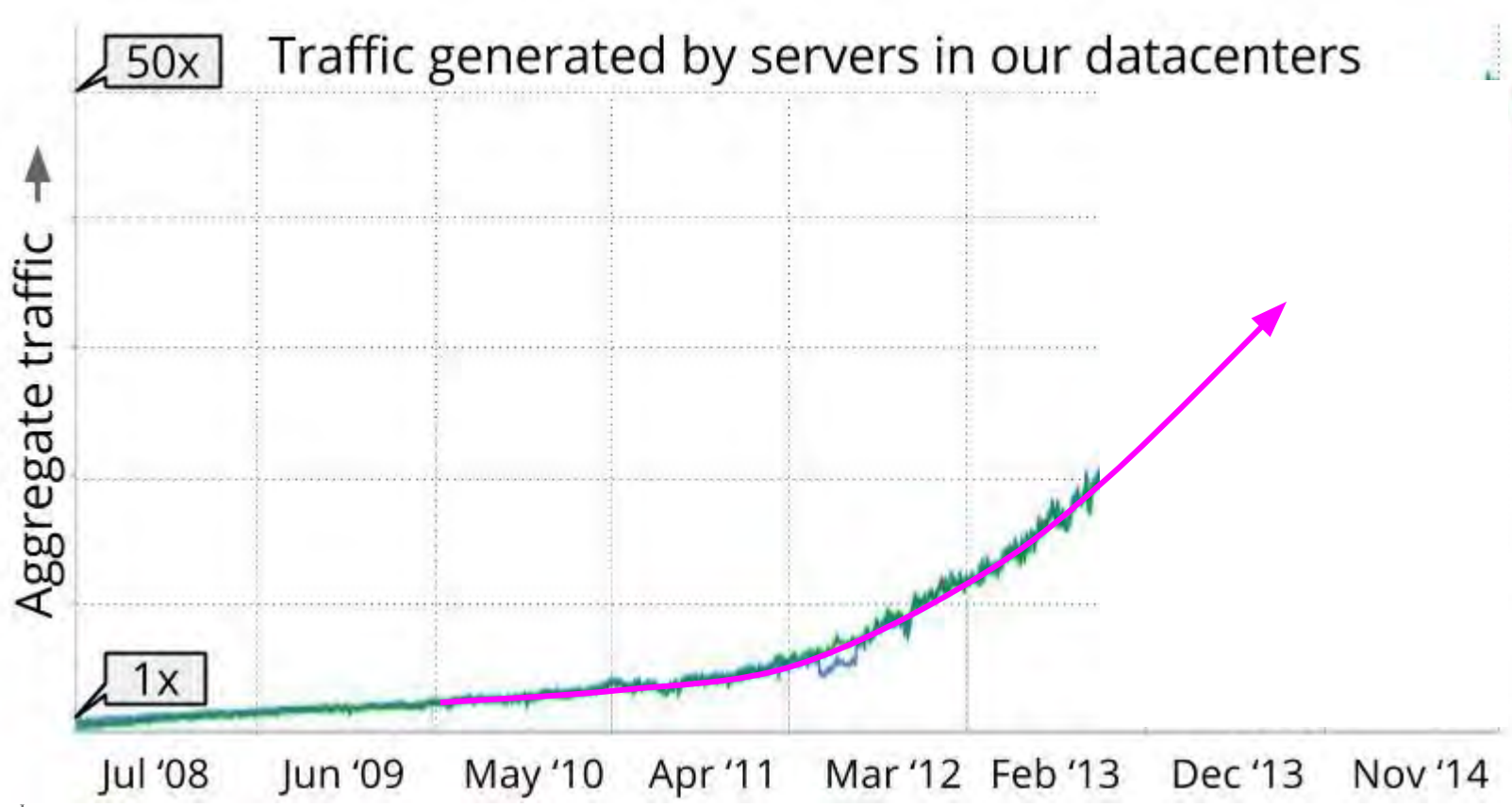
$$ prices



compute/storage capacity

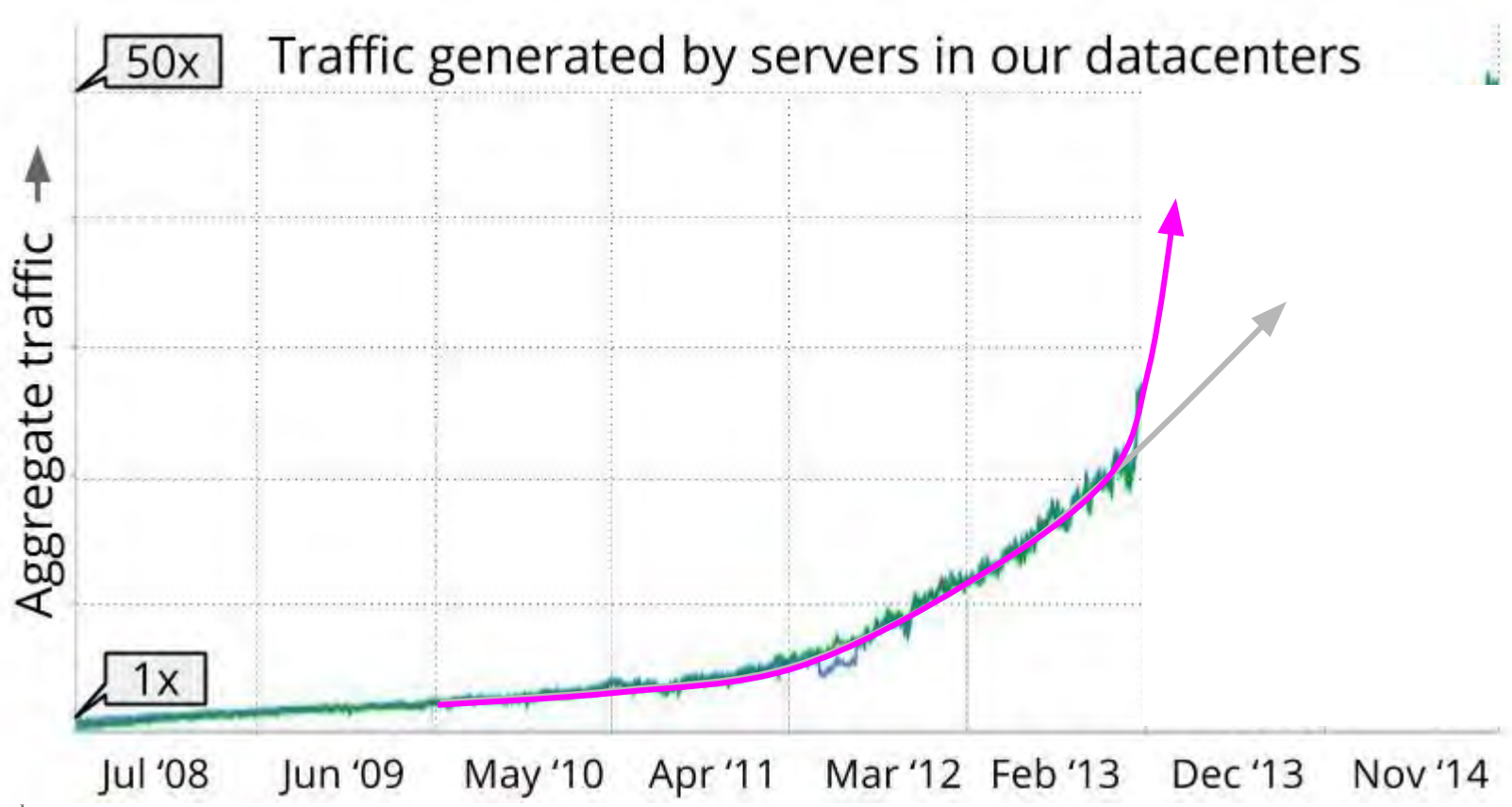
Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network. SIGCOMM 2015.

In a world of exponential demand growth



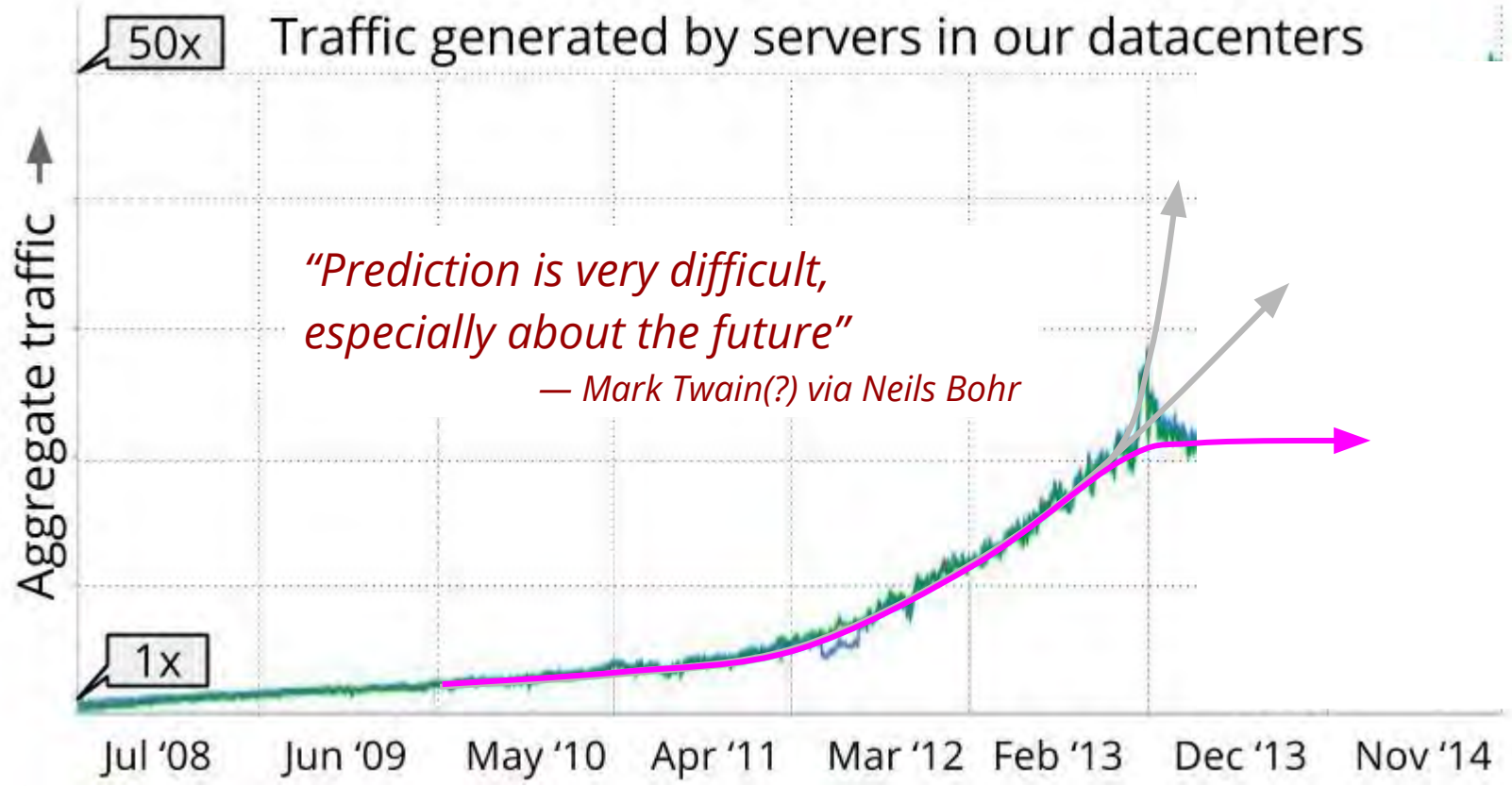
Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network. SIGCOMM 2015.

In a world of exponential demand growth

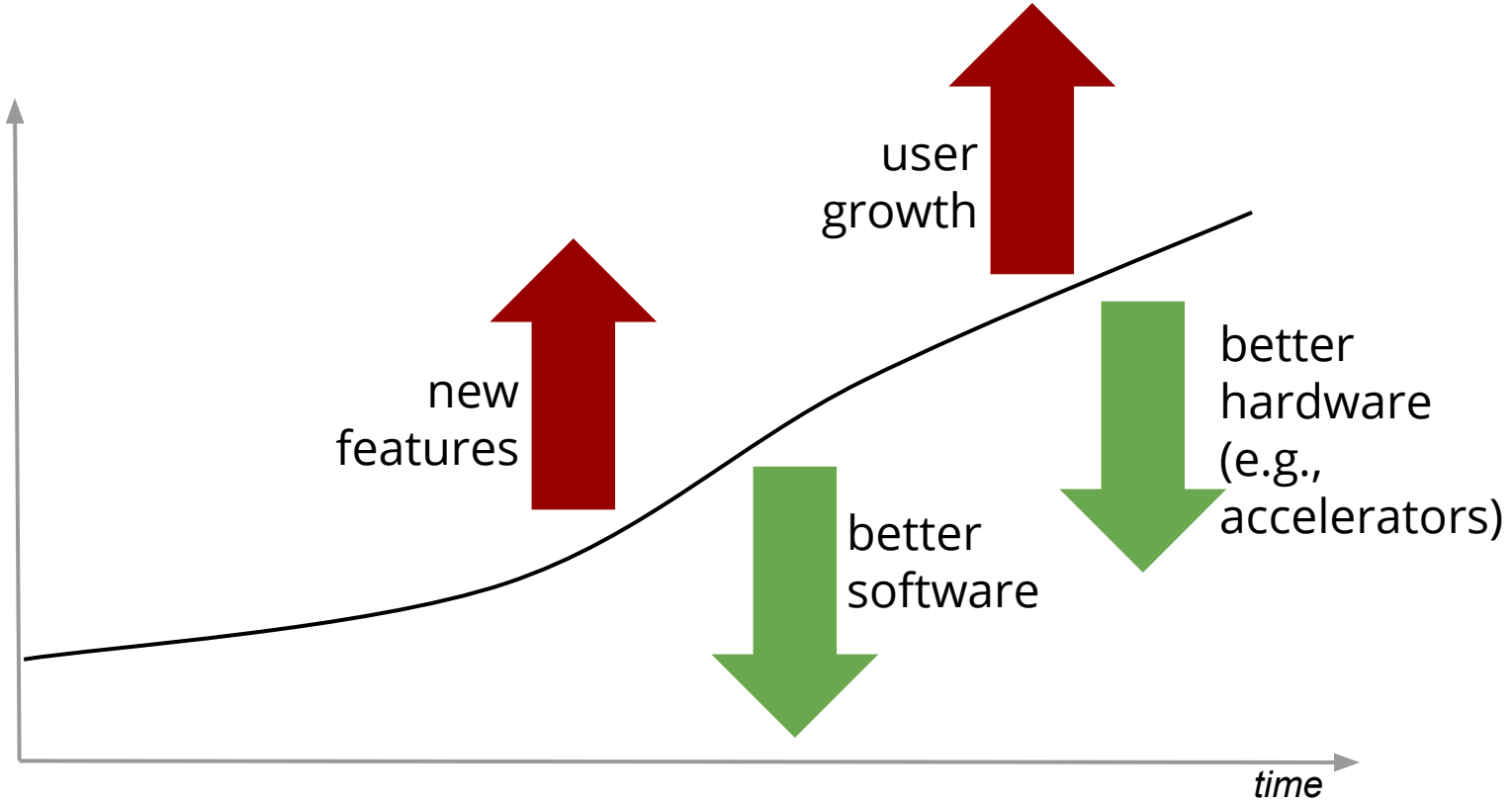


Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network. SIGCOMM 2015.

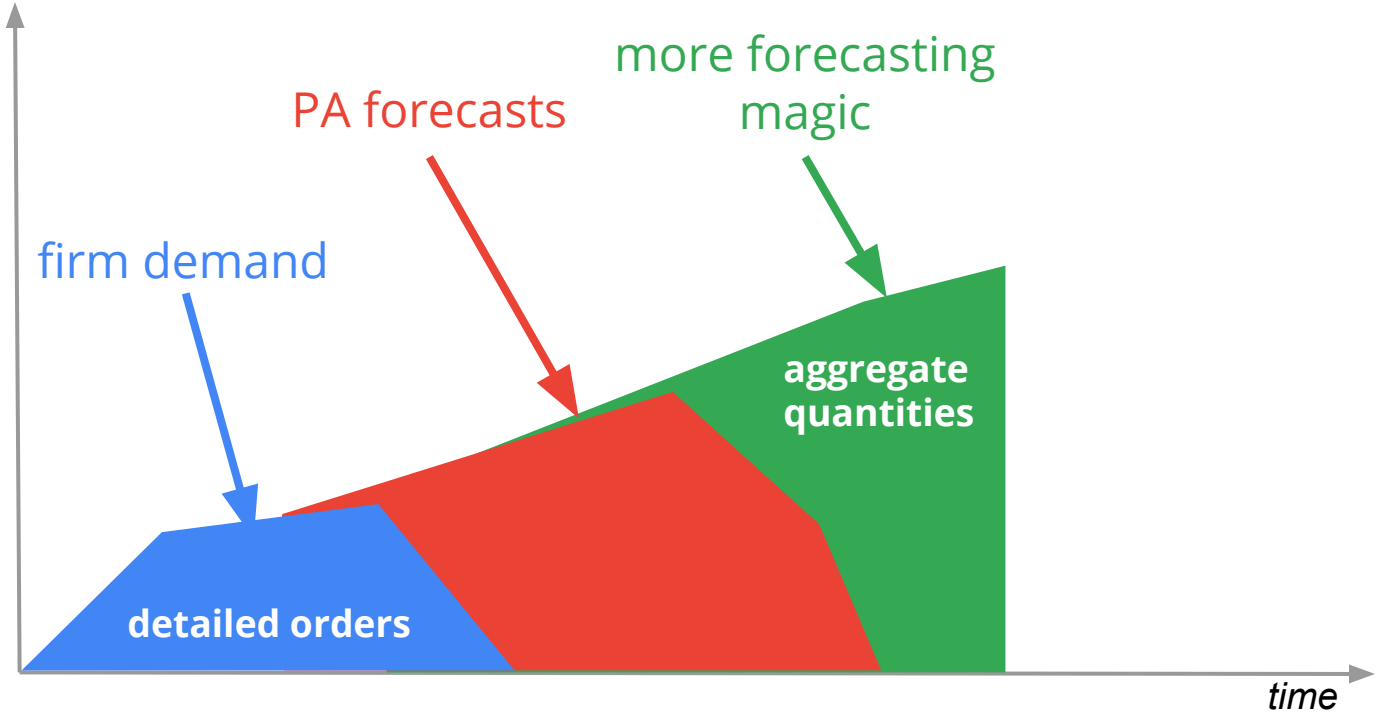
In a world of exponential demand growth



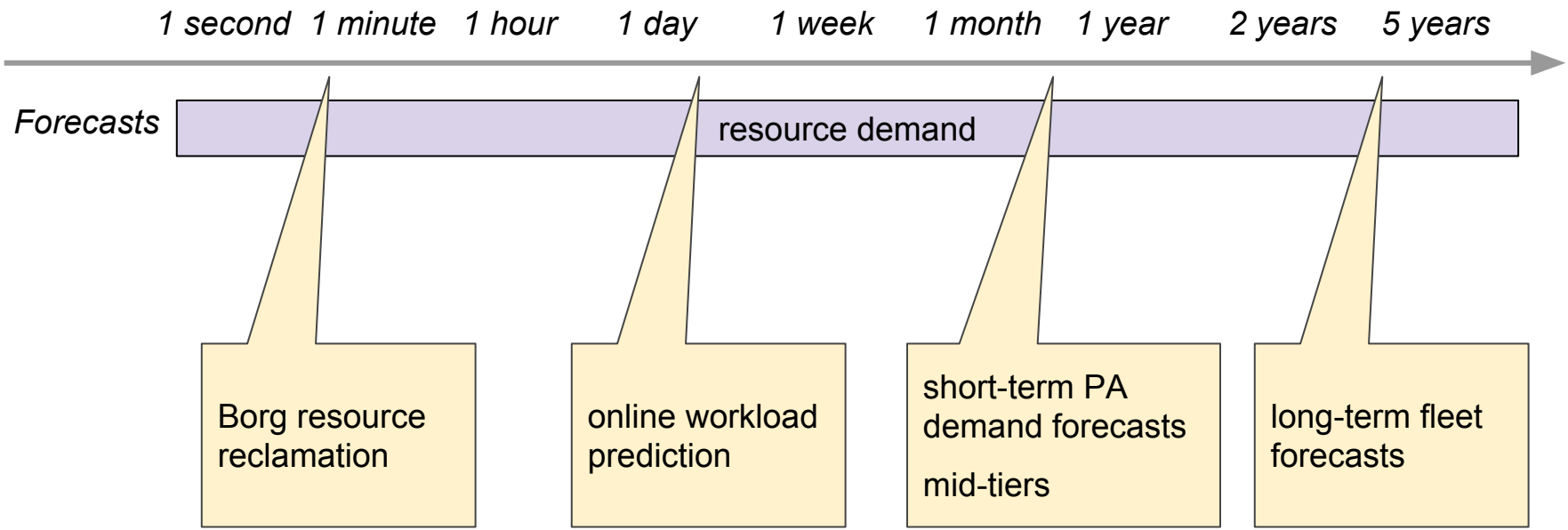
A few factors affecting forecasts



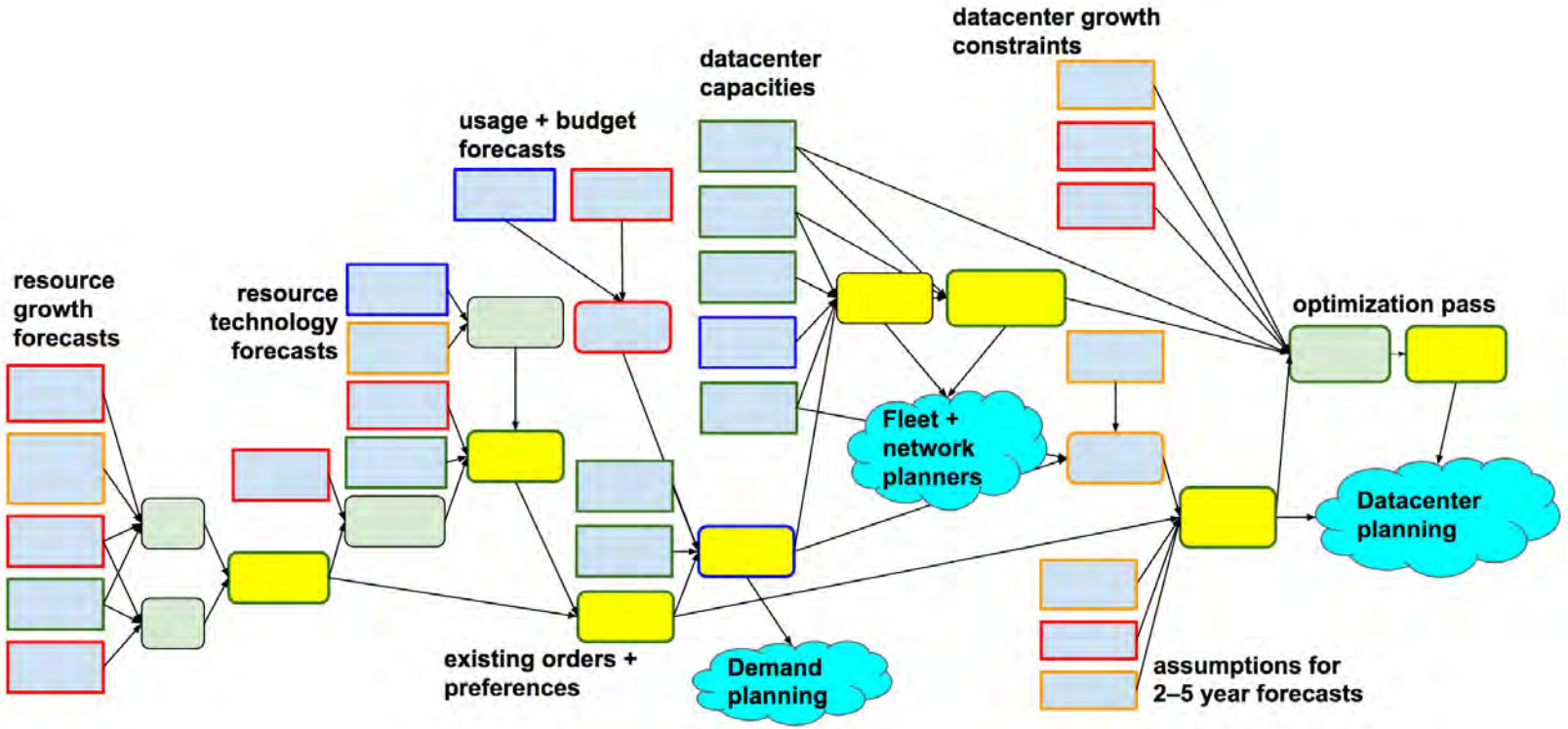
How much capacity do we need?
and when do you need to know?



Putting it all together



It takes a few moving parts ...

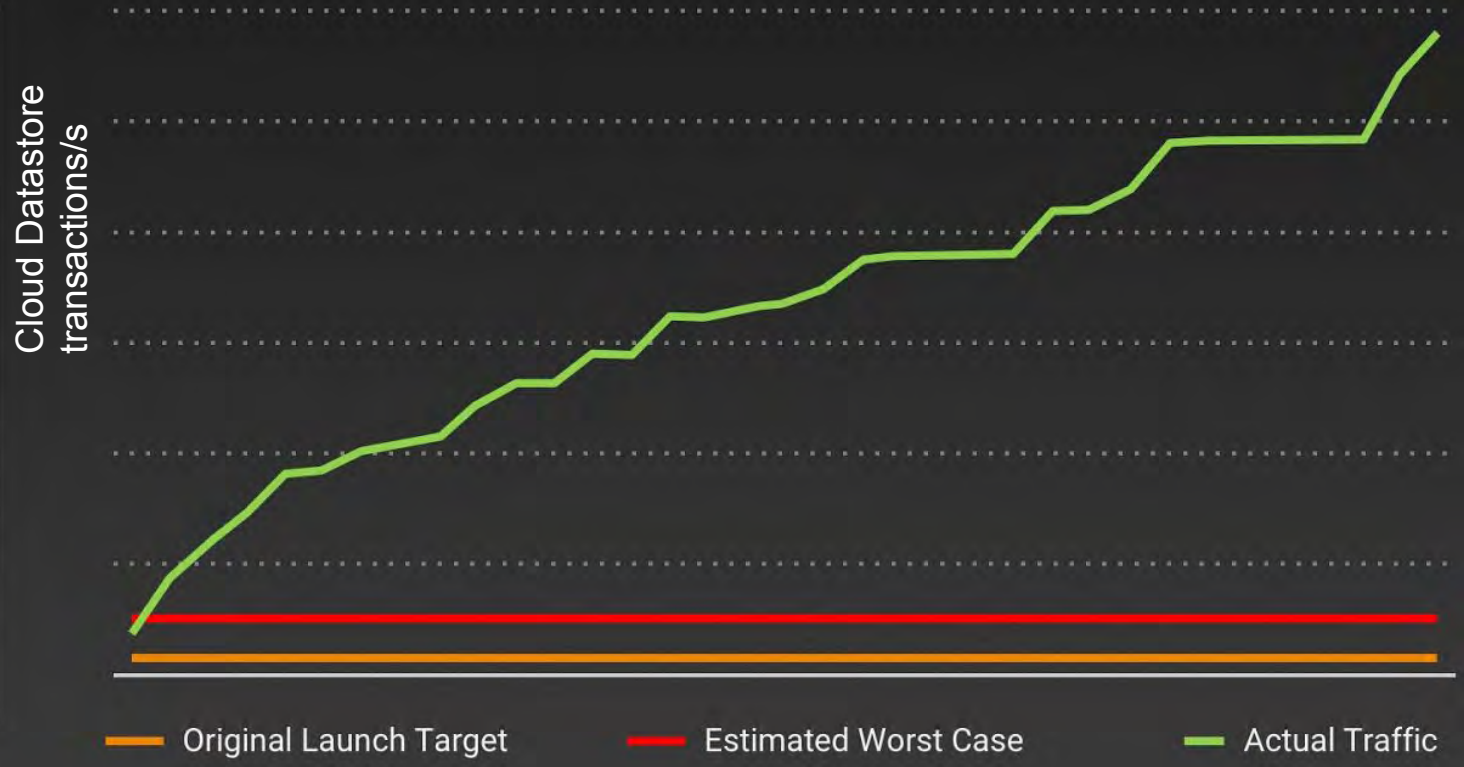


[redacted]

1X
Target Traffic

5X
Worst Case Estimate

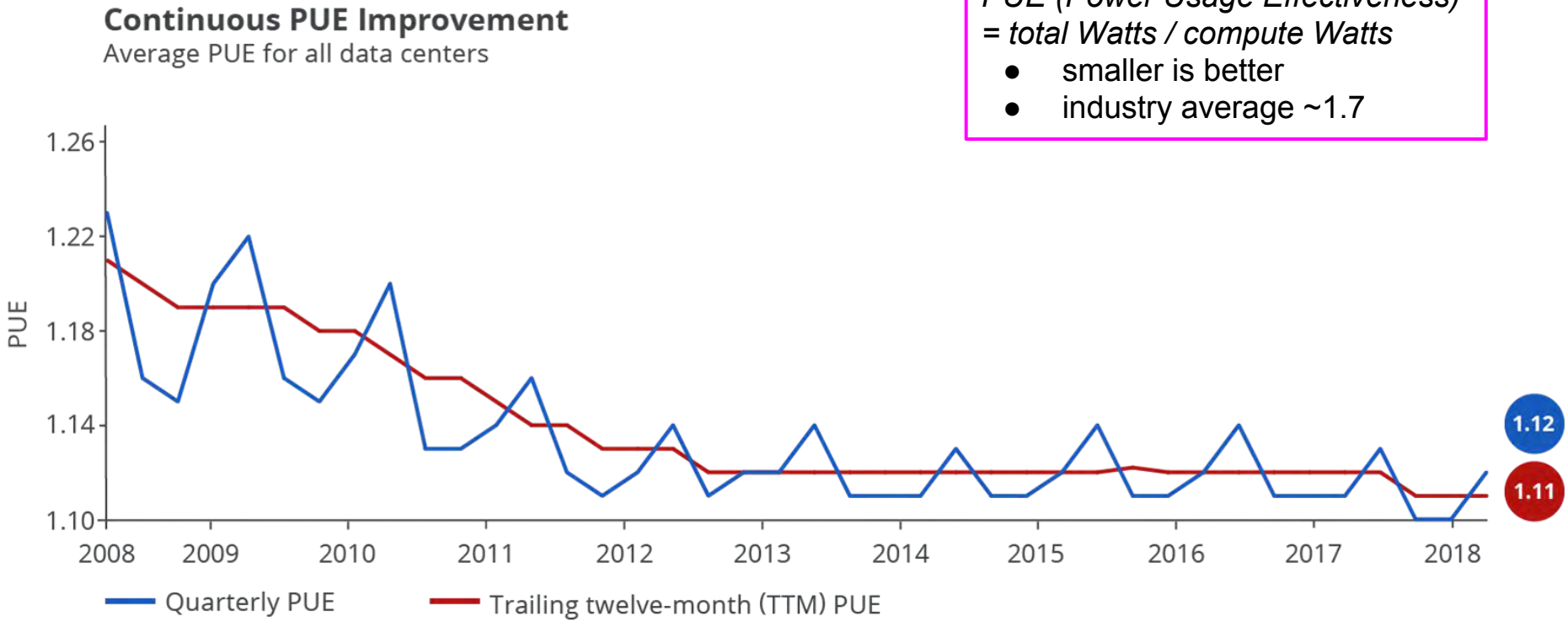
50X
Actual Traffic



Planning for power

PUE (Power Usage Effectiveness)
= total Watts / compute Watts

- smaller is better
- industry average ~1.7



https://www.google.com/about/datacenters/efficiency/internal/

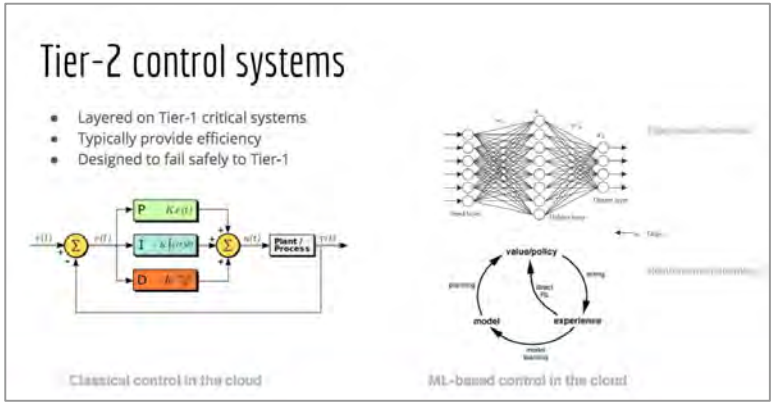
Planning for power

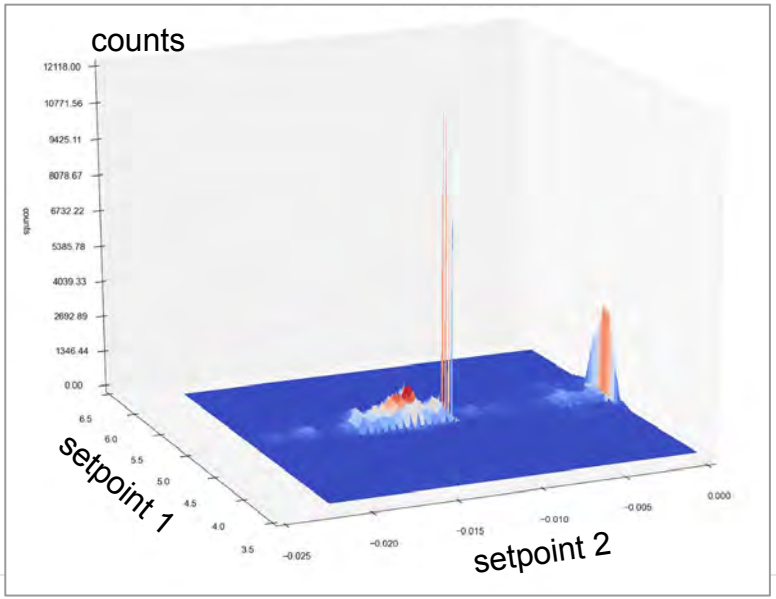
Planning for power

Some of the challenges

- Narrow range of experience
- Exploration may discover unsafe states
- Inputs out of our control (e.g., weather)
- Control system reliability/availability
- Agility (new hardware)
- Reinforcement learning with long delays between action and change in system state

- **Safety. Safety. Safety.**





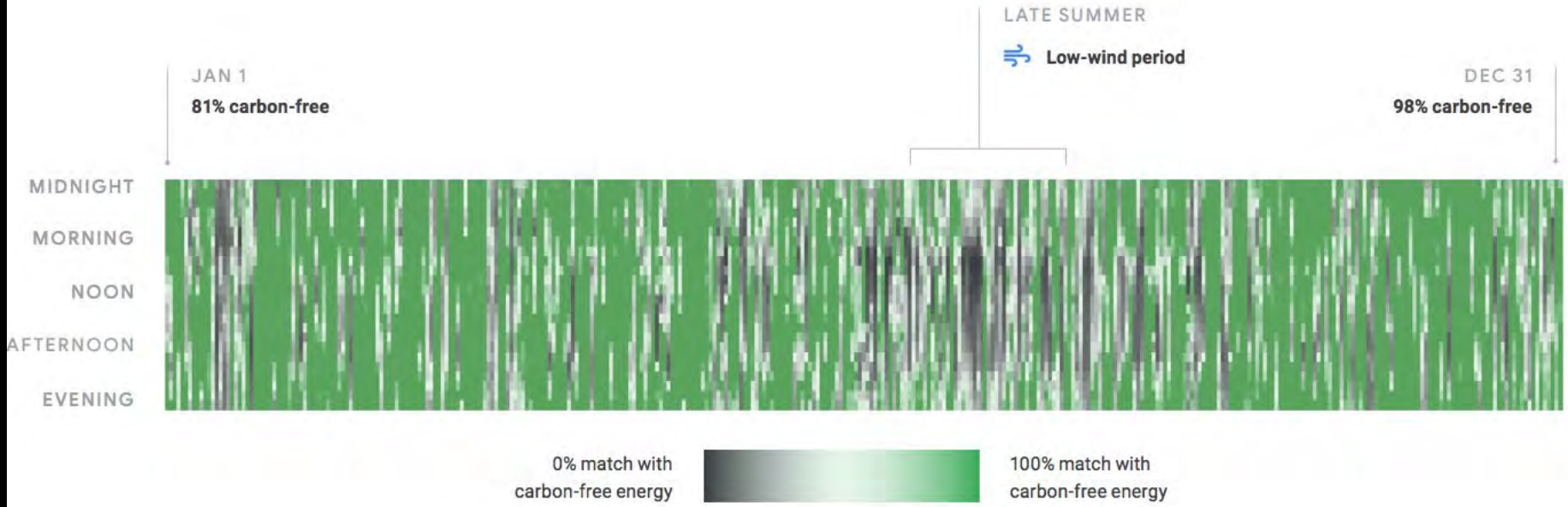
Graphics by Steve Webster


We’re set to reach 100% renewable energy in 2017

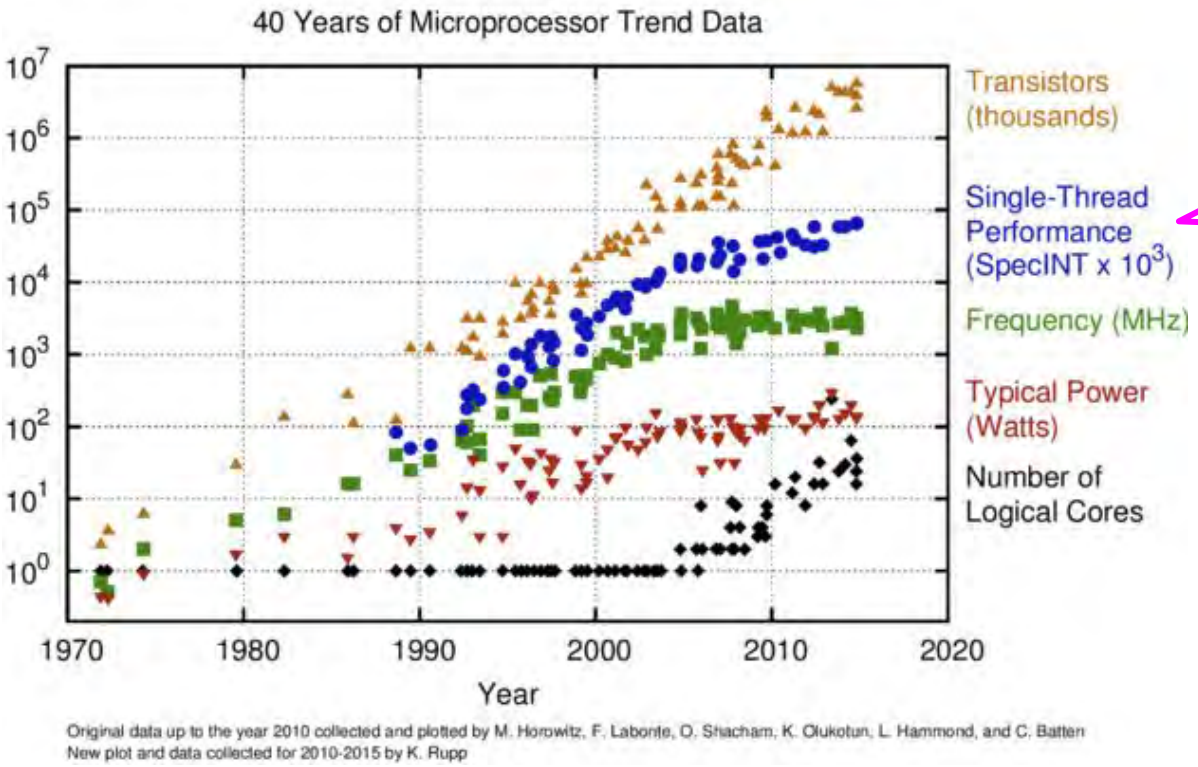
Every hour of electricity use at Iowa data center

Although our Iowa data center achieved 100% carbon-free energy during the majority of hours in 2017, there is also a recurring reliance on carbon-based power — most notably in late summer, when wind speeds decline.

Overall in 2017, 74% of this data center's electricity use was matched on an hourly basis with carbon-free sources.



Meanwhile – what's up with Moore's law?

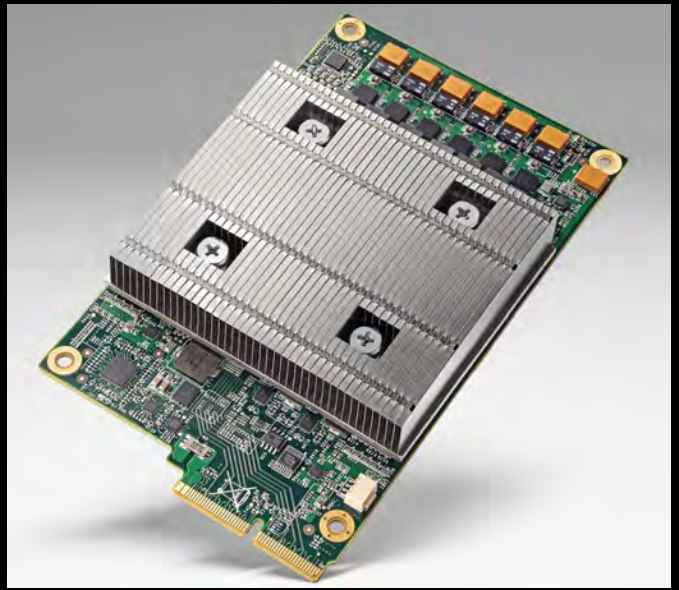


Single-core performance plateauing after decades of exponential growth

Graph from 40 Years of Microprocessor Trend Data, Karl Rupp, CC-BY 4.0.

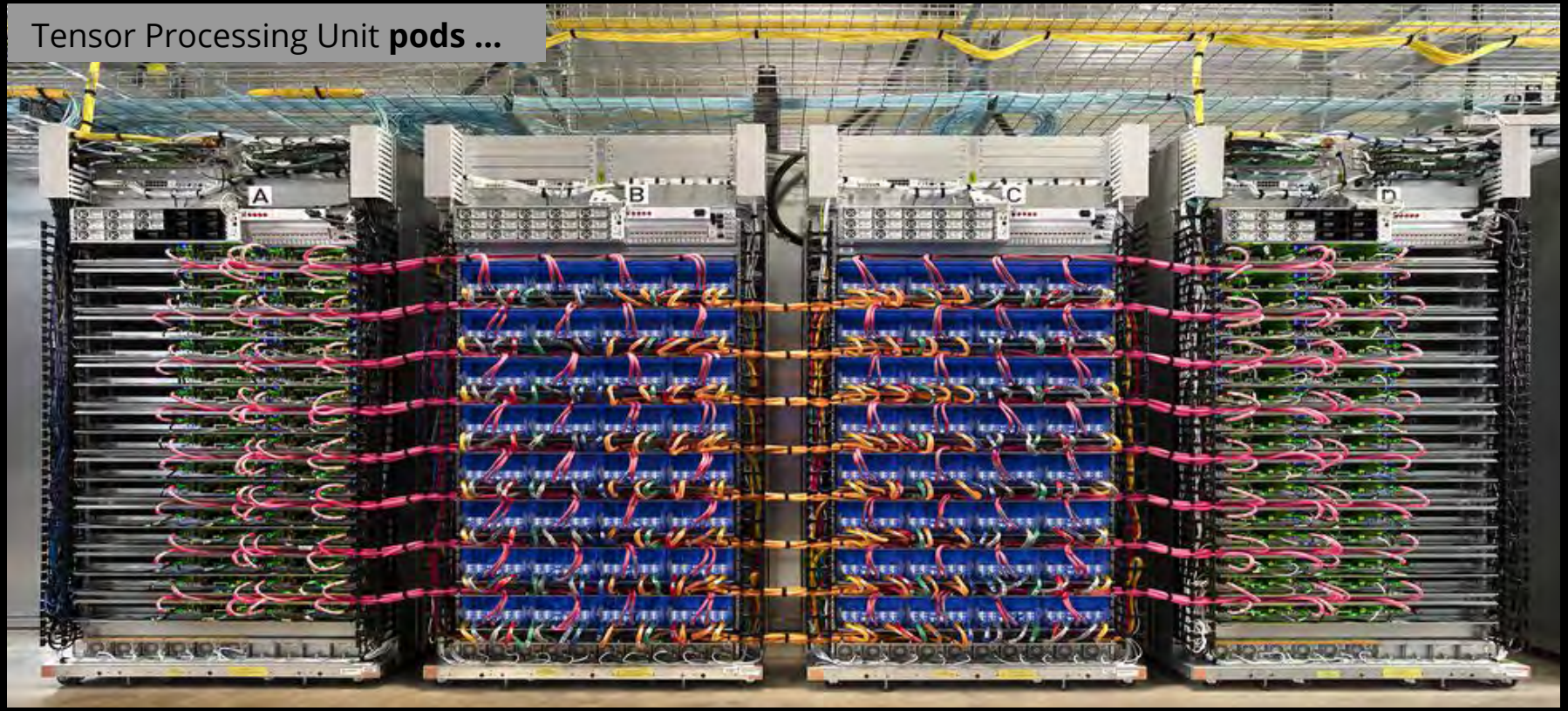
Meanwhile – what's up with Moore's law?

Tensor Processing Units

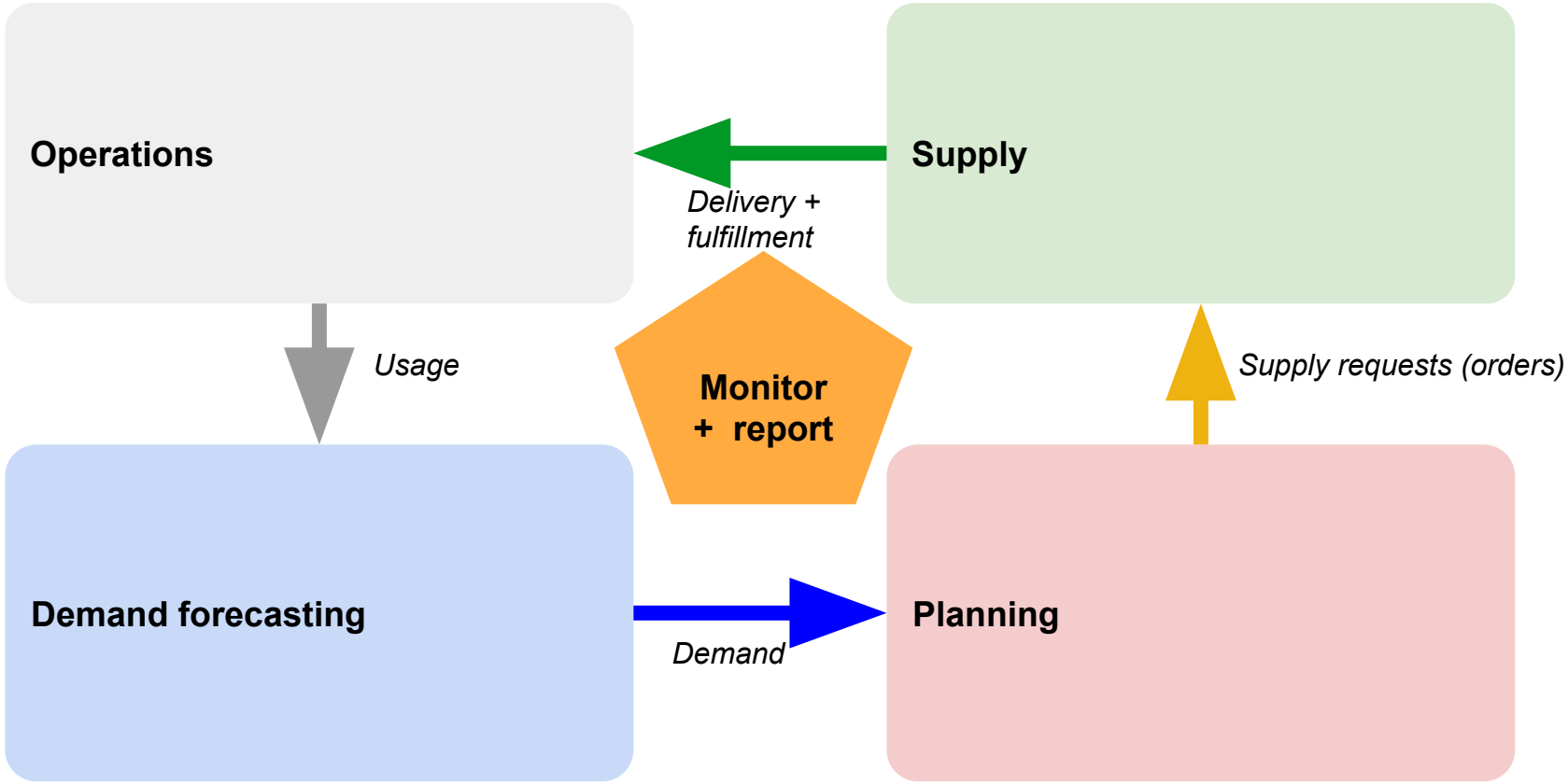


Meanwhile – what's up with Moore's law?

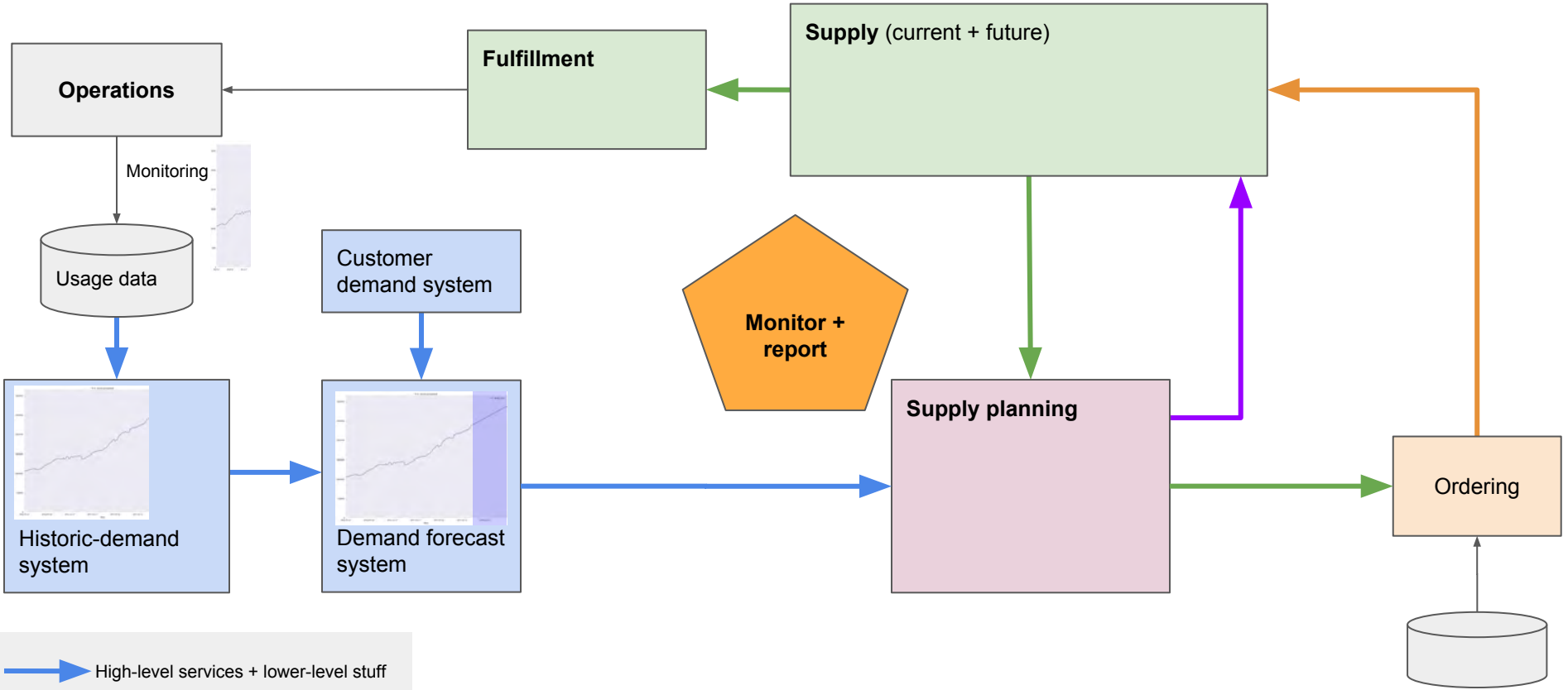
Tensor Processing Unit **pods …**



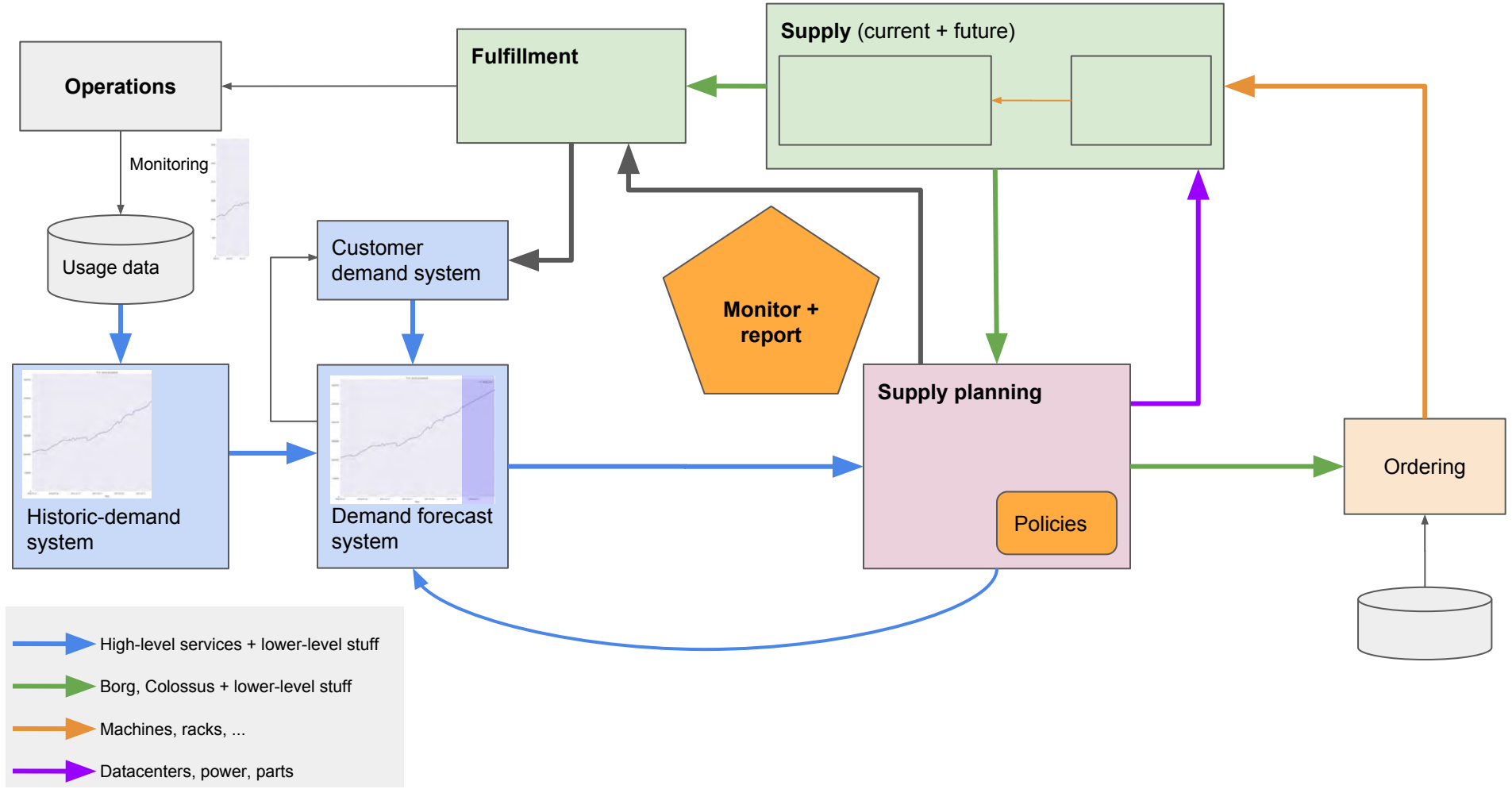
Putting it all together



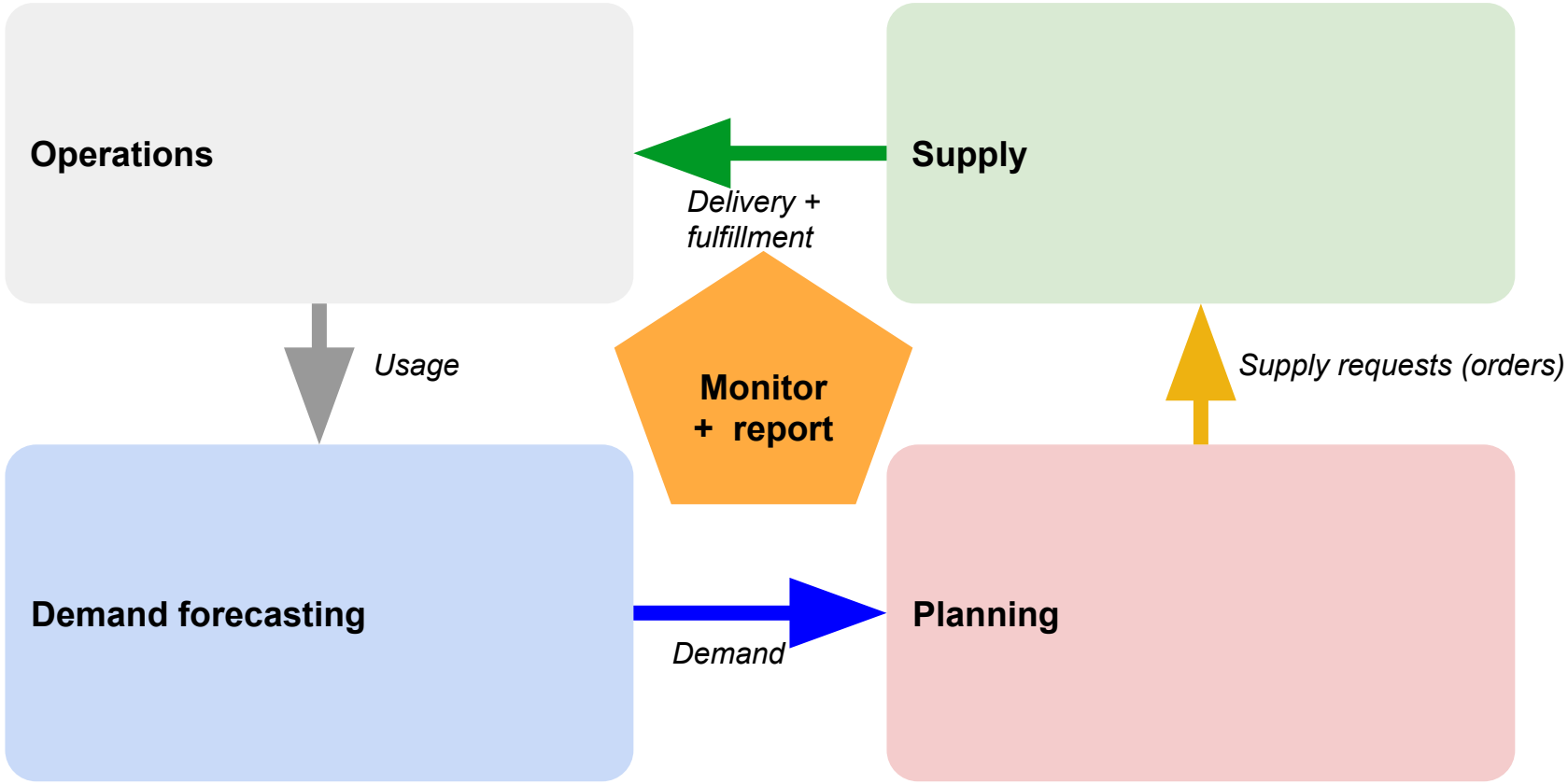
Putting it all together – a few more details



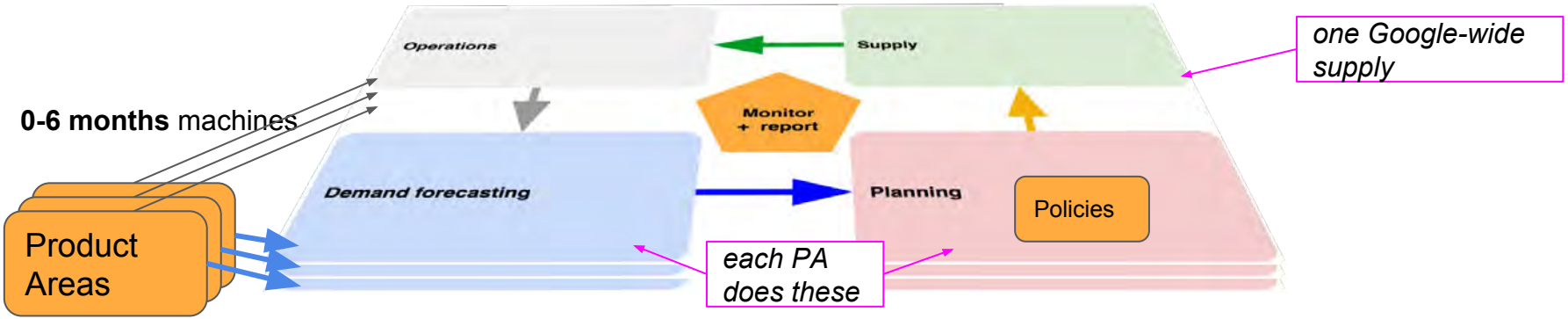
Putting it all together – a few more details



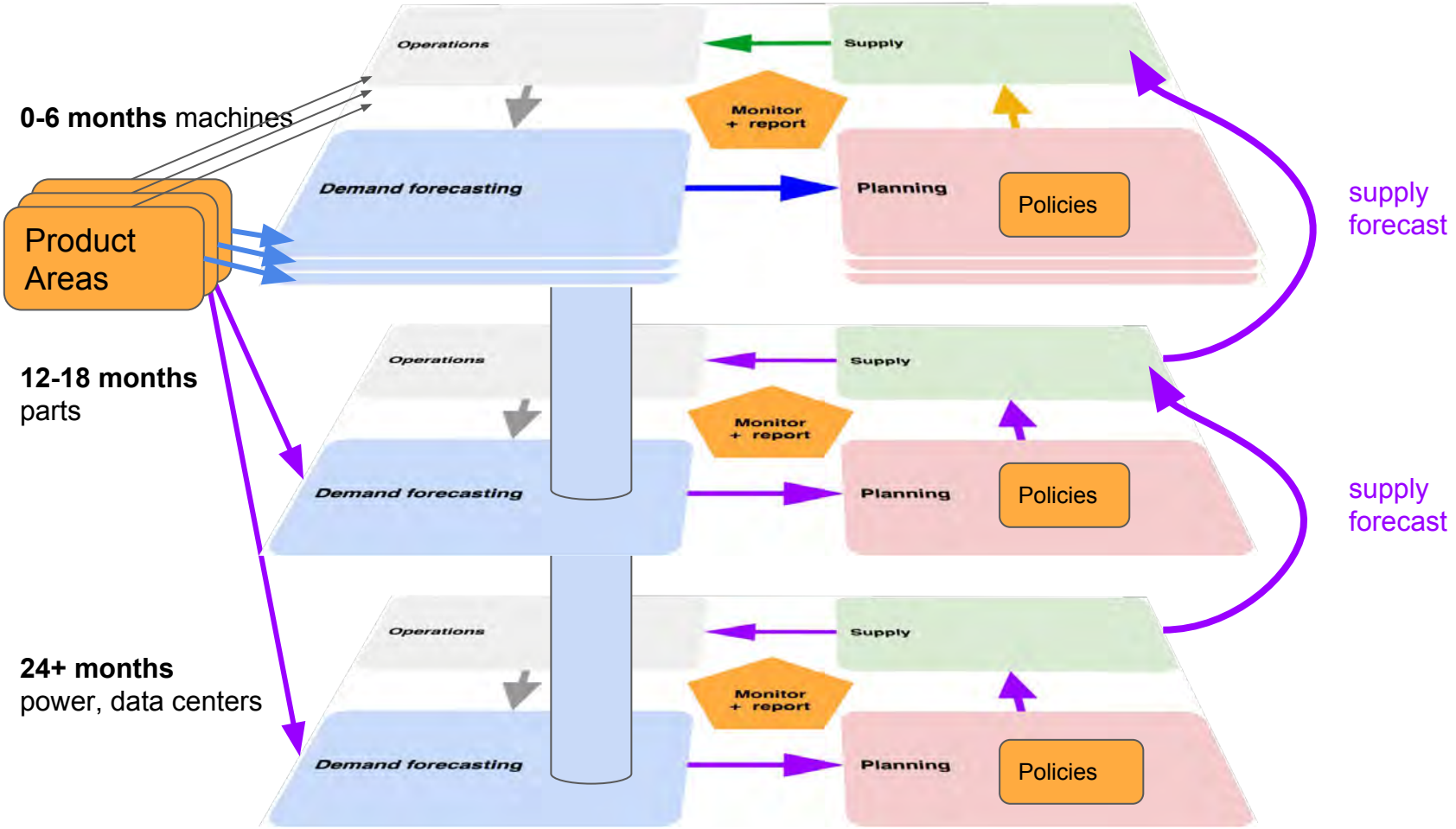
Putting it all together



Putting it all together



Putting it all together – multiple timelines




Hello world!
Image by Connie Zhou

2018 Q1 CapEx = $5.3B

(+$2.4B for an office building in New York)

source: Alphabet SEC filing

$29.4B

3-year trailing CapEx, as of March 2017

Final thoughts:

There's a lot of technology behind "the cloud"

At scale, efficiency *really* matters

[and: we're hiring!]

johnwilkes@google.com

Images by Connie Zhou