Things to read while you wait:

- **Creating an effective poster**, john wilkes
  http://goo.gl/HUVffd

- **The 4 sentence abstract**, Kent Beck
  https://goo.gl/W5SCMT

Building warehouse-scale computers

or … what's it like to supply exponential growth

john wilkes 2018-11

Google Cloud Platform

Video: https://youtu.be/m7uIG8qFGMI

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.

~~You're all~~

Some of you are

thinking

too small

Google Cloud Platform | Google Cluster Data

BigQuery BETA | Go to Classic UI | + COMPOSE NEW QUERY

Query history
Saved queries
Job history
Transfers
Resources + ADD DATA
Search for your tables and datasets
google.com:google-cluster-data
google.com:bigquery-petabyte

2018-11-05 1PB scan | LINK SHARING | HIDE EDITOR

```
SELECT *
FROM `google.com:bigquery-petabyte.retail_petabyte.sales_partitioned`
WHERE customerKey = "1104796800000-19548818"
```

Run query | Save query | Save view | More

Query results

Query complete (5 min 49.309 sec elapsed, 1.09 PB processed)

Job information | Results | JSON | Execution details

For help debugging or optimizing your query, check our documentation. Learn more

Elapsed time	Slot time consumed	Bytes shuffled	Bytes spilled to disk
5 min 49.309 sec	5 days 16 hr	0 B	0 B

Worker timing

Stages		Wait	Read	Compute	Write		Rows
S00: Input	Avg:	149717 ms	387 ms	45 ms	22 ms	Input:	1,006,914,092,880
	Max:	309819 ms	44565 ms	9260 ms	17947 ms	Output:	0

Show build data

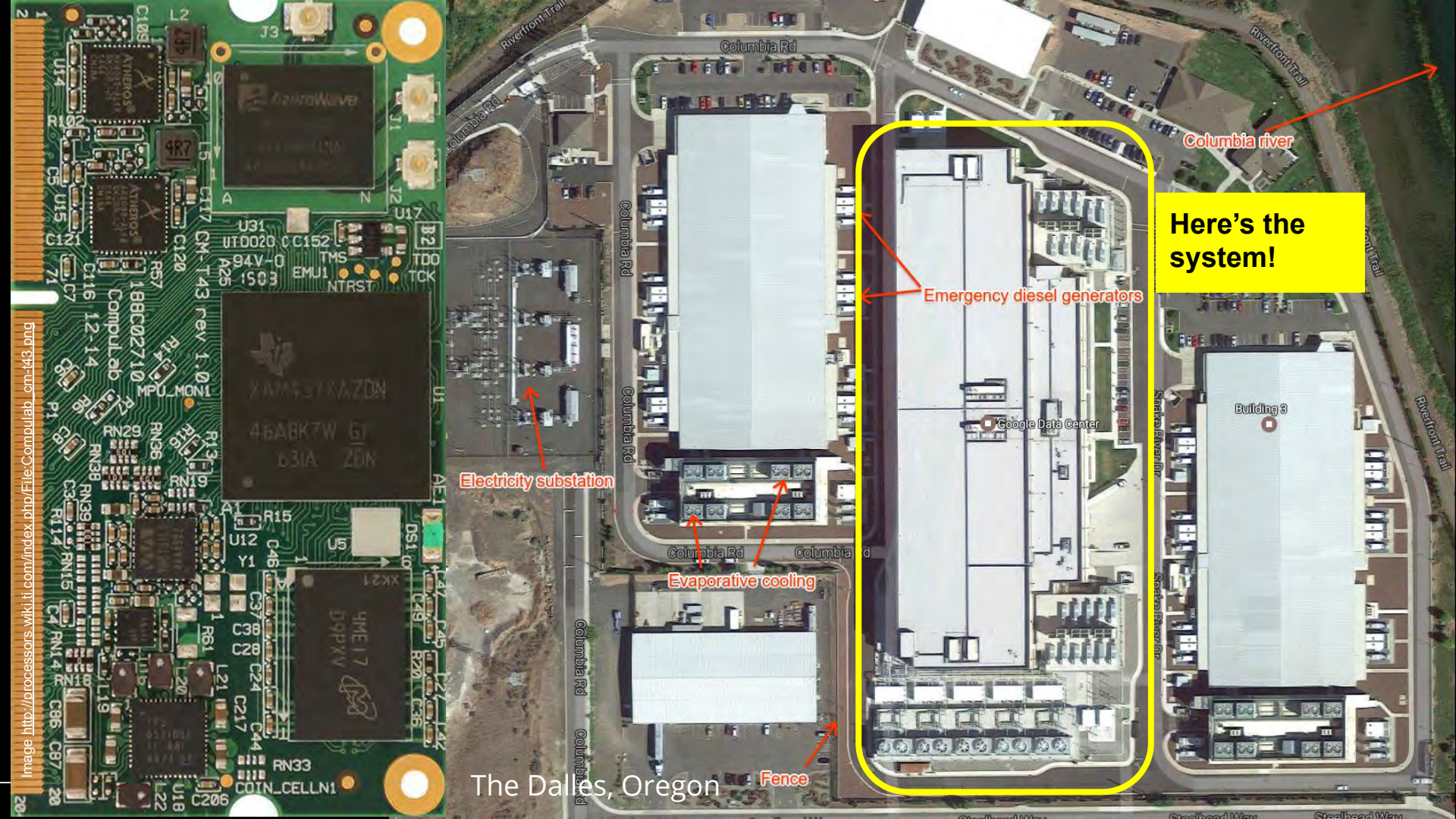


Image http://processors.wiki.ti.com/index.php/File:Compulab_cm-t43.png

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

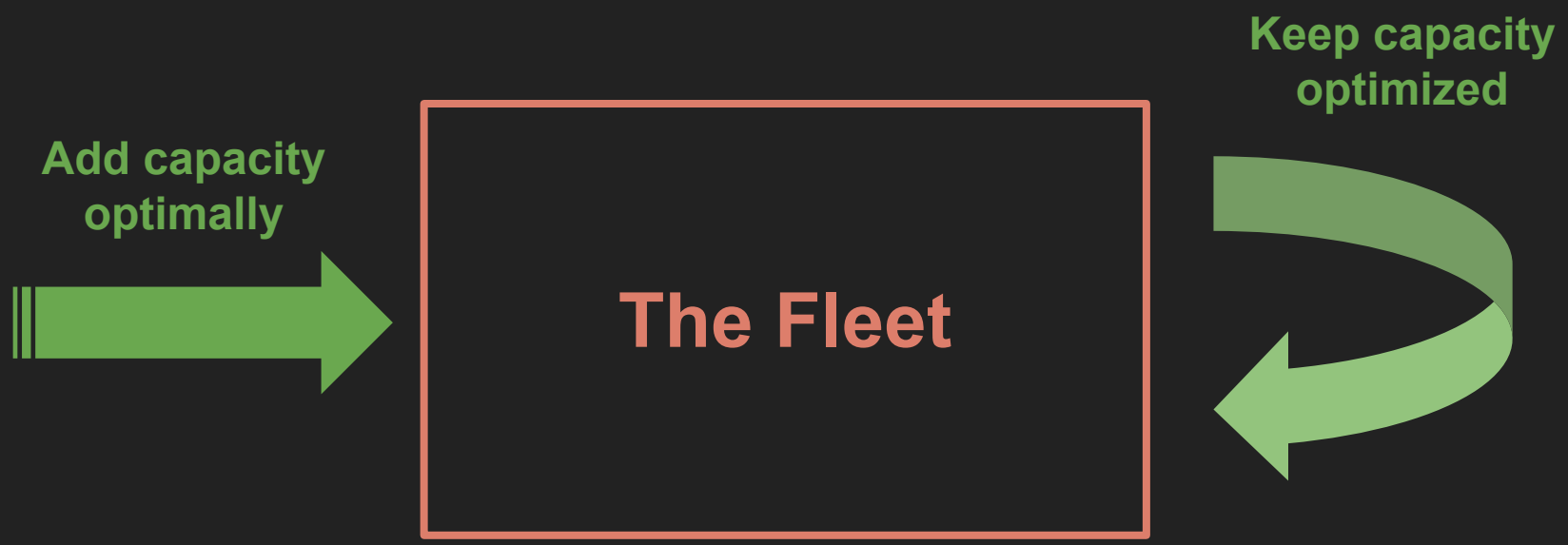
and the power system
and the technicians

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

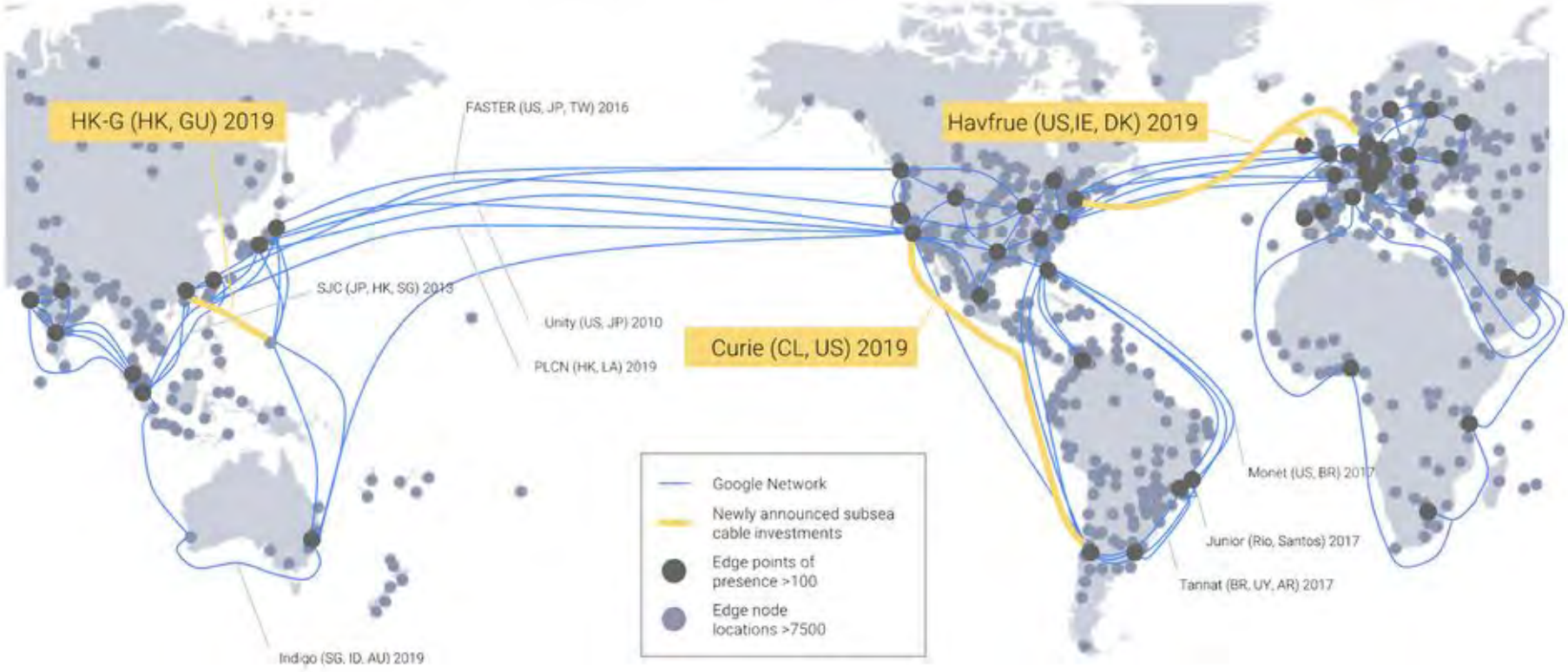
- new NVRAM thingies …
- accelerators (complication: a wide variety)
- inter- and intra-datacenter networking
- power
- datacenters
- land, water, sewage
- …

Response times are weeks/months/years, not milliseconds

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

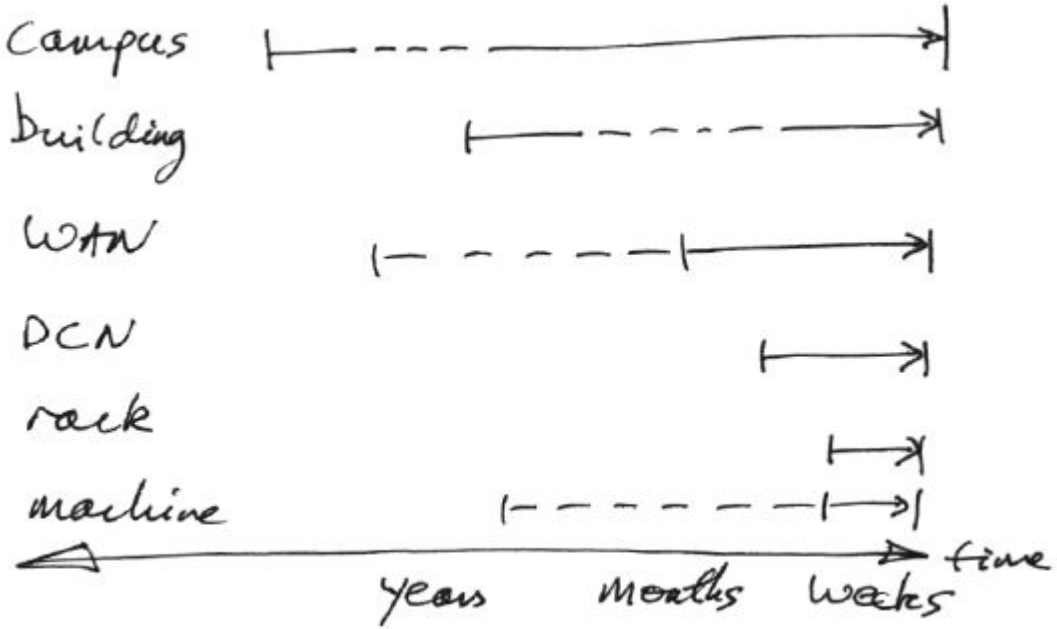
Google Network

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**"?

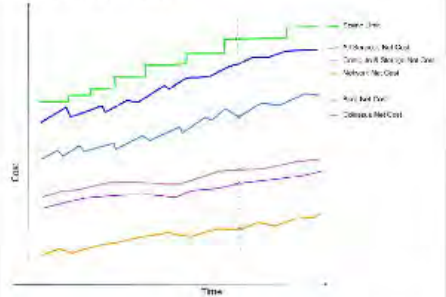
- machines can be transferred/given/sold/break
- who "owns" a machine running a shared service for a customer?

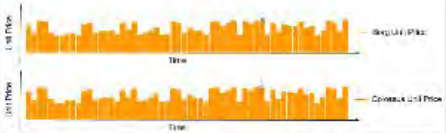
Resource Trends

org "adwords serving" location:world resource (all computestorage network borg colossus)

Cost Trends — Unit Price — 2 Metrics

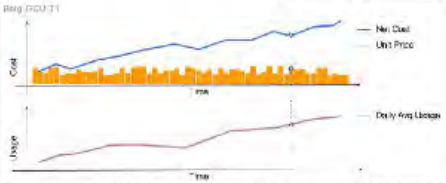
Organization: Serving (PG) ▸ — Location: World ▸





Resource	Net Cost	Unit Price
All Services	[redacted]	--
Compute & Storage	[redacted]	--
Network	[redacted]	--
Borg	[redacted]	[redacted]
Colossus	[redacted]	[redacted]

Product Trends — 2 Products — 3 Metrics



Net Cost	Unit Price	Daily Avg Usage
[redacted]	[redacted]	[redacted]

Inventory

group:it-flex-pool location:region=asia service:bigtabl

Summary

Group	Daily Cost	Health
ti-flex-pool ▸	[redacted]	Healthy
Location	**Footprints**	**Scaling**
Region Asia ▸	[redacted]	1 needs attention ▸
Service	**Registration Status**	2 executing ▸
Bigtable ▸	Open	10 recent ▸

Product Metrics — 4 Metrics

Product	Capacity	Commitment	Available	Available %
Servers	554	499	55	10%
GCU	[redacted]	[redacted]	[redacted]	28%
RAM	[redacted]	[redacted]	[redacted]	61%
Low Latency	[redacted]	[redacted]	[redacted]	10%
Throughput	[redacted]	[redacted]	[redacted]	28%
Disk	[redacted]	[redacted]	[redacted]	61%
SSD	[redacted]	[redacted]	[redacted]	10%

Capacity Breakdown — Disk

Metric Timelines — Total Cost

Fleet Optimization Study Execution

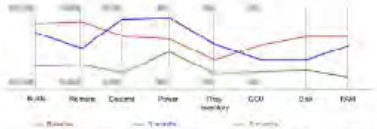
id: 5678

Summary

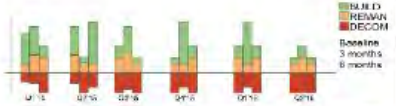
ID	Executed On
5678	2016-01-01 12:00 UTC
Base Fleet Optimization	**Status**
123456 ▸	Success
Study Name	**Study ID**
Haswell Delay	1234 ▸

Description: [redacted]

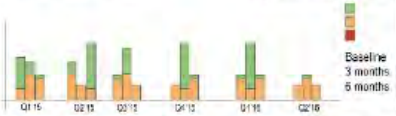
Comparison Summary



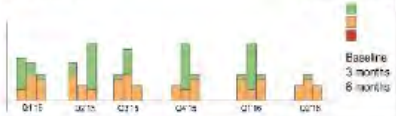
Compare Machines — world=world



Compare Hub Inventory — 3 Hubs — 1 Tray



Compare Power — world=world



Order

id:123454321

Summary

ID	Supplier
12345432	machine-service-so
Service	**Consumer**
PA-Specified Machine	ads-pantry
Type	**Purpose**
Create	Product launch
Location	**Category**
Cluster io ▸	None
Requested start date	**Relationship**
2016-12-31	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

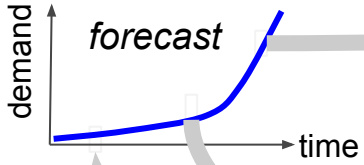
Requested	Ordered	Planned	Delivered
GTIM00003 - [redacted]			
3k	3k	--	--
GTIM00004 - [redacted]			
4k	4k	--	--
GTIM00005 - [redacted]			
4k	4k	--	--

Open Forecast for Lego Inventory

Cost — Monthly Breakdown

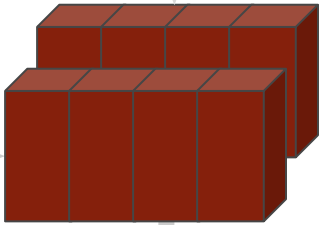
	Ordered	Planned	Delivered	Pending Delivery
Dec 2016	[redacted]	[redacted]	[redacted]	[redacted]
Jan 2017	[redacted]	[redacted]	[redacted]	[redacted]
Feb 2017	[redacted]	[redacted]	[redacted]	[redacted]
Mar 2017	[redacted]	[redacted]	[redacted]	[redacted]
Apr 2017	[redacted]	[redacted]	[redacted]	[redacted]

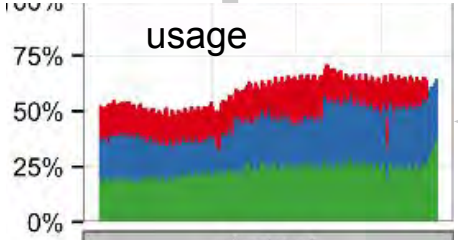
A simplified overview



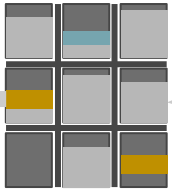
sites, data centers, power

network, machines, storage





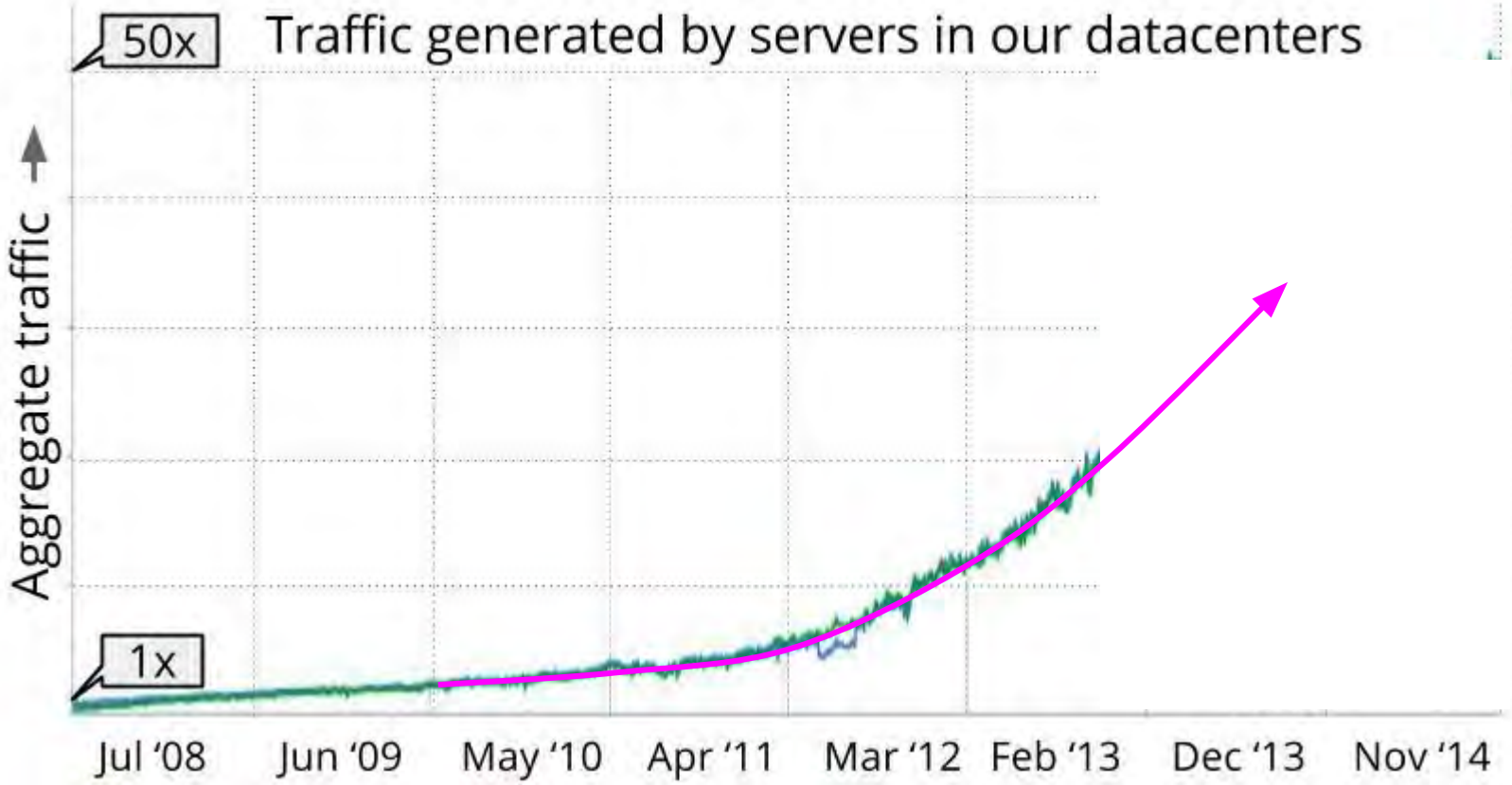
$$ prices



compute/storage capacity

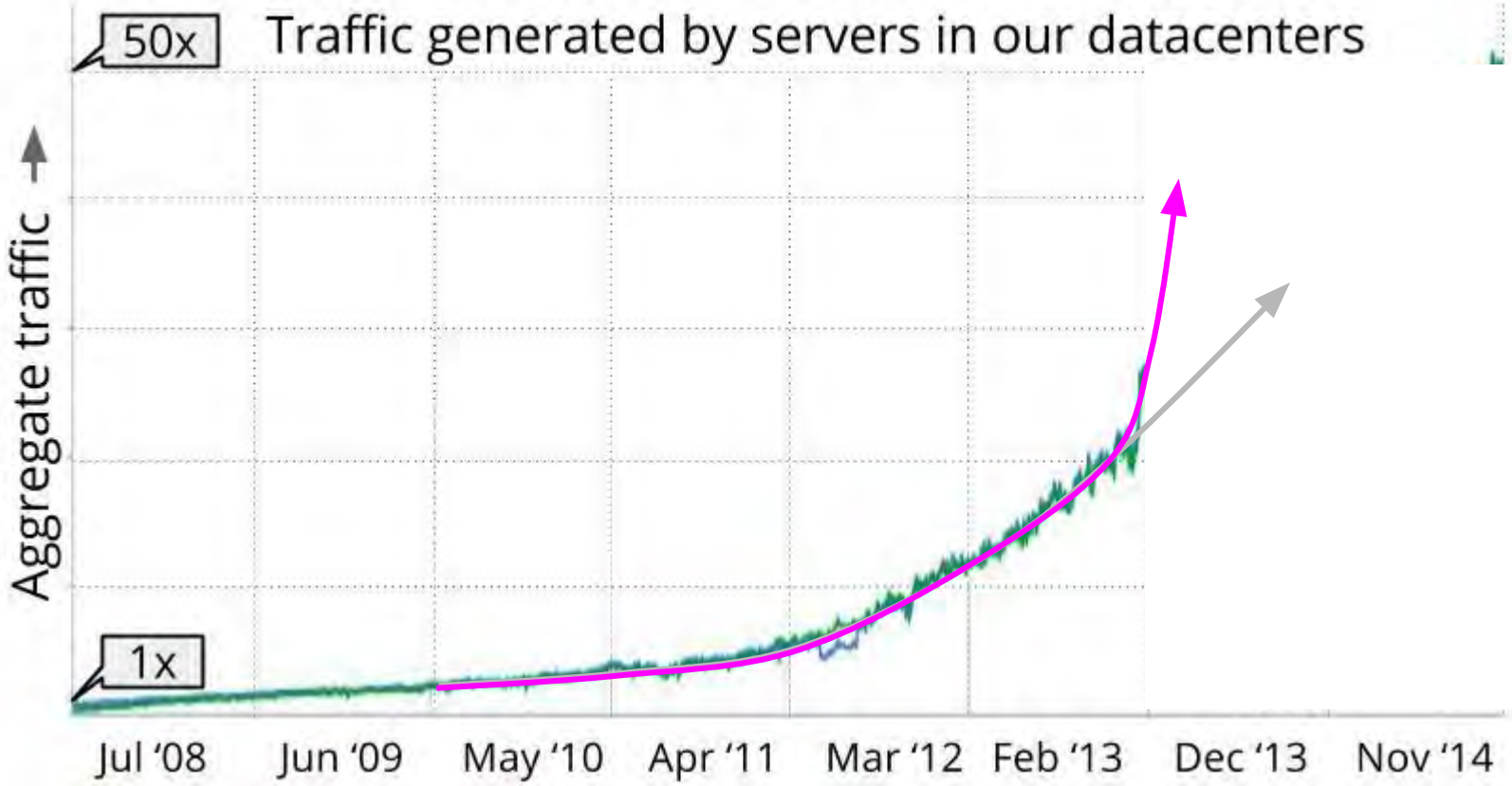
Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network. SIGCOMM 2015.

In a world of exponential demand growth



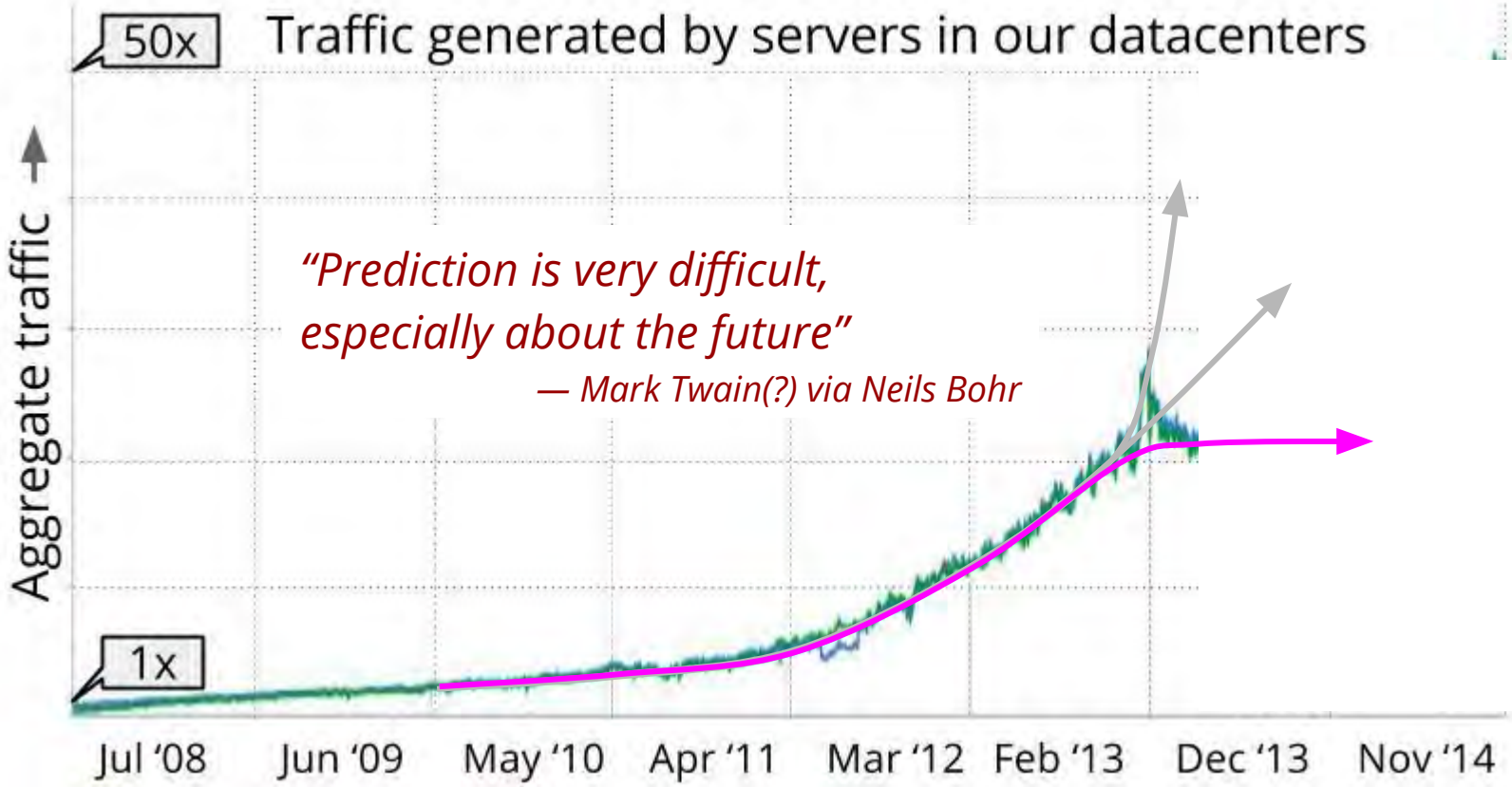
Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network. SIGCOMM 2015.

In a world of exponential demand growth



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

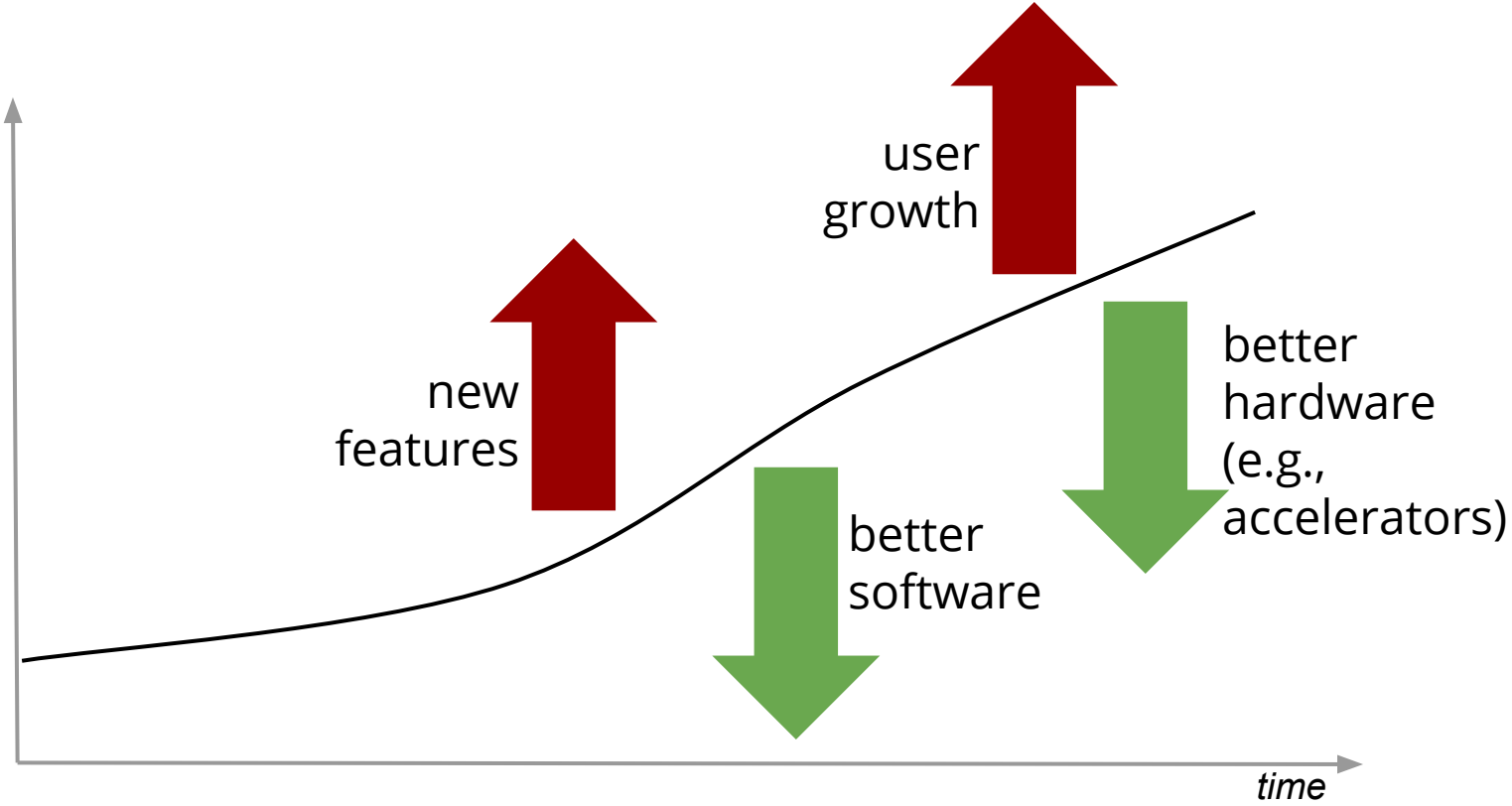
In a world of exponential demand growth



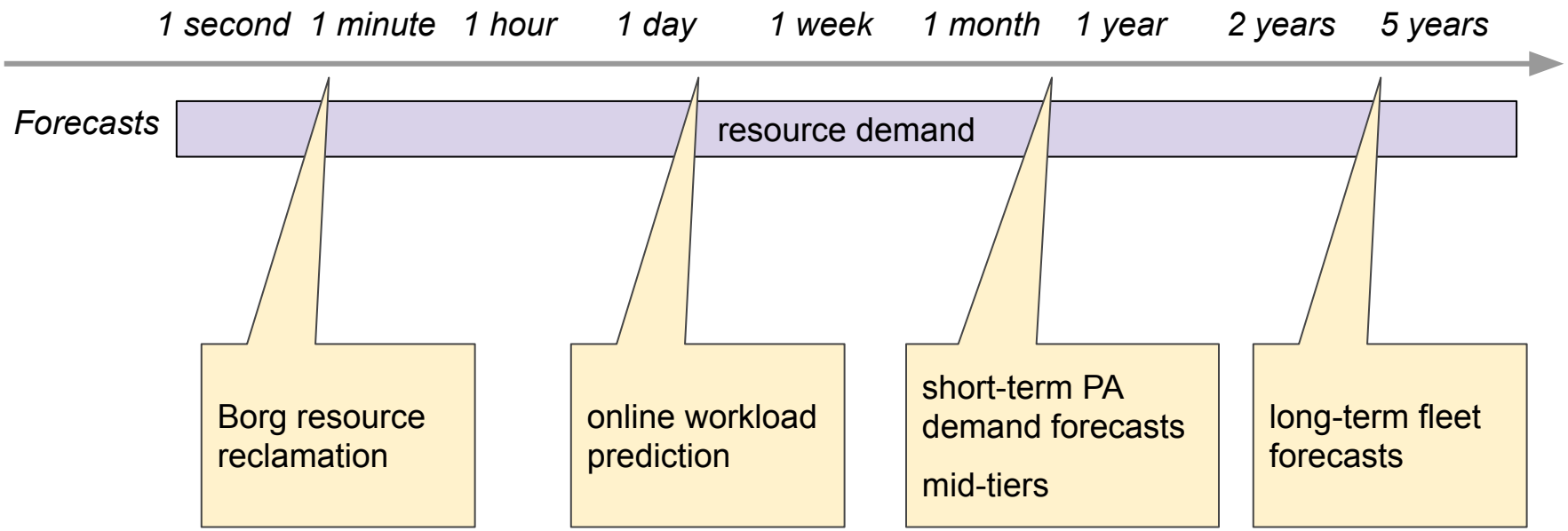
"Prediction is very difficult, especially about the future"

— Mark Twain(?) via Neils Bohr

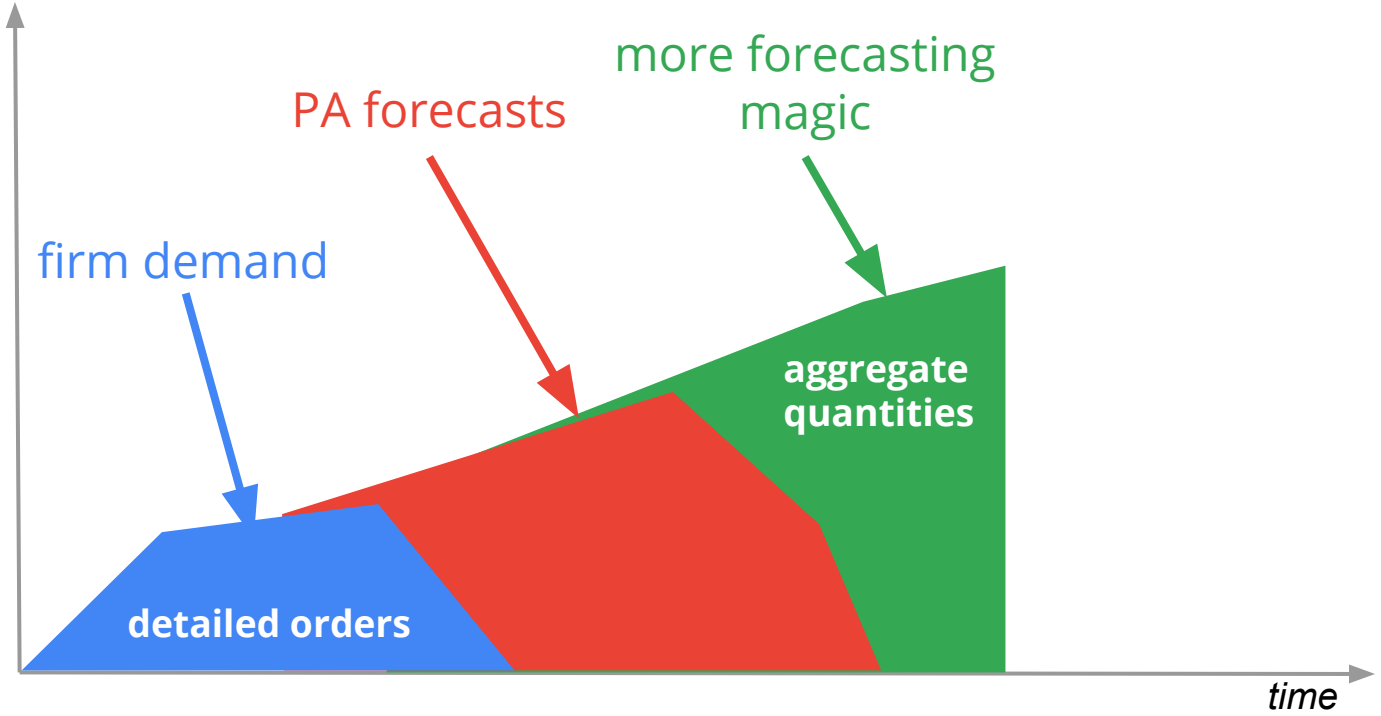
A few factors affecting forecasts



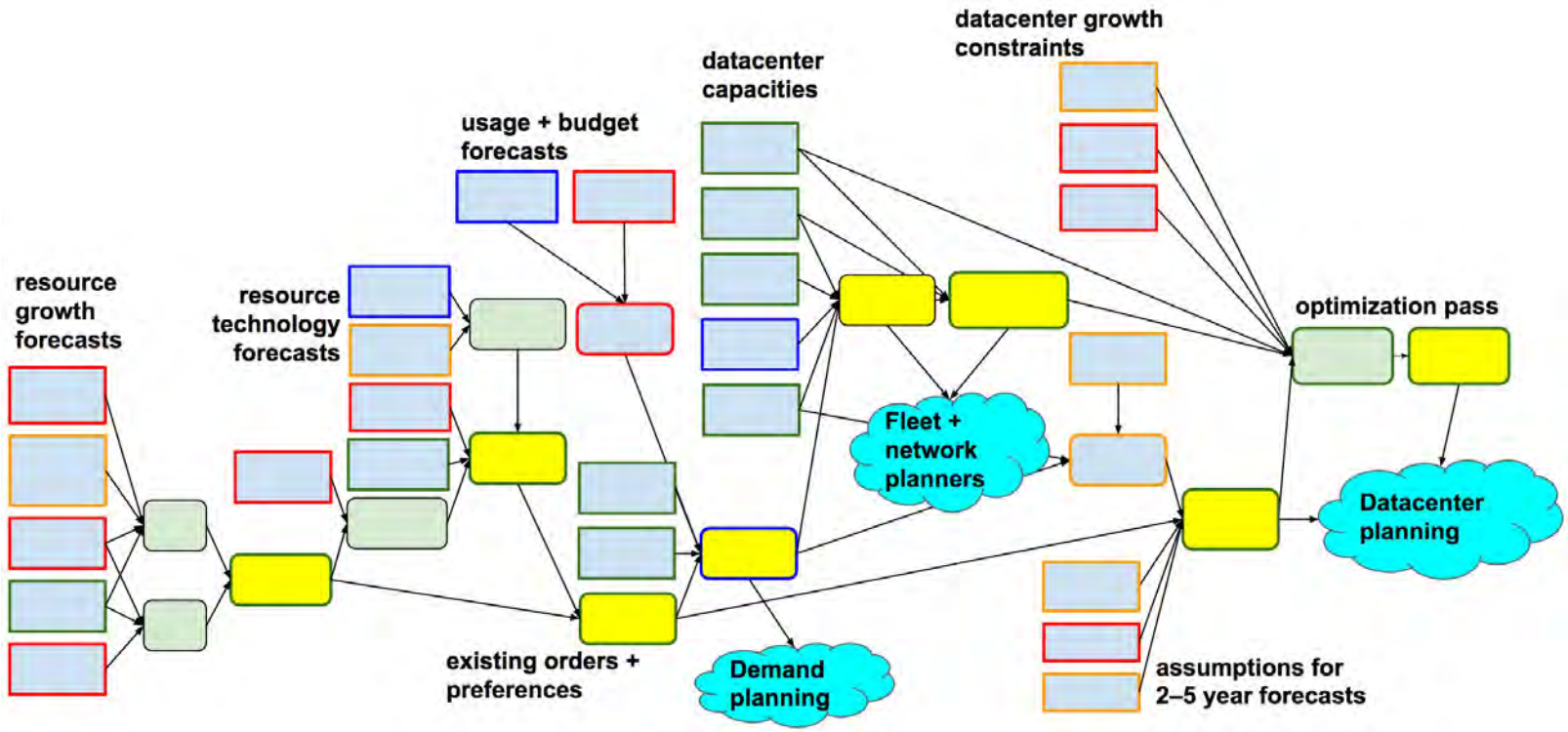
Putting it all together



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



It takes a few moving parts ...



[redacted]

1X
Target Traffic

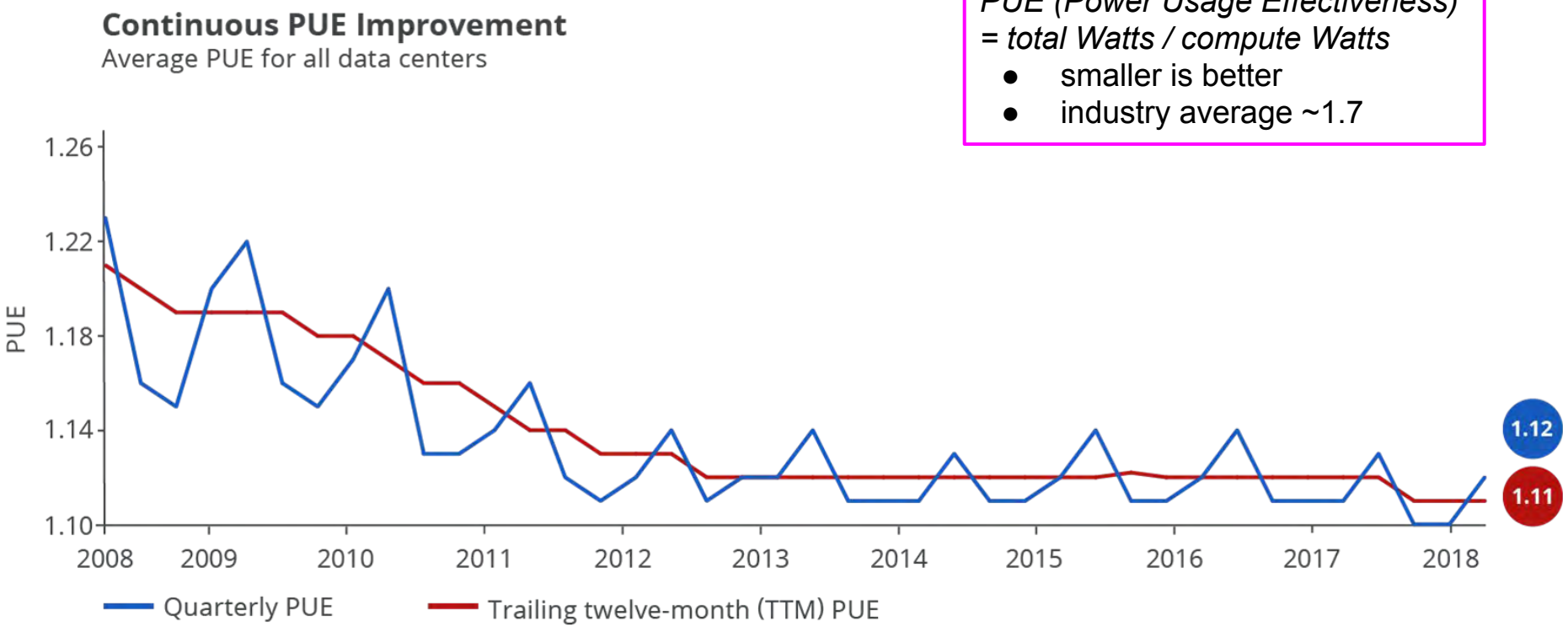
5X
Worst Case Estimate

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/

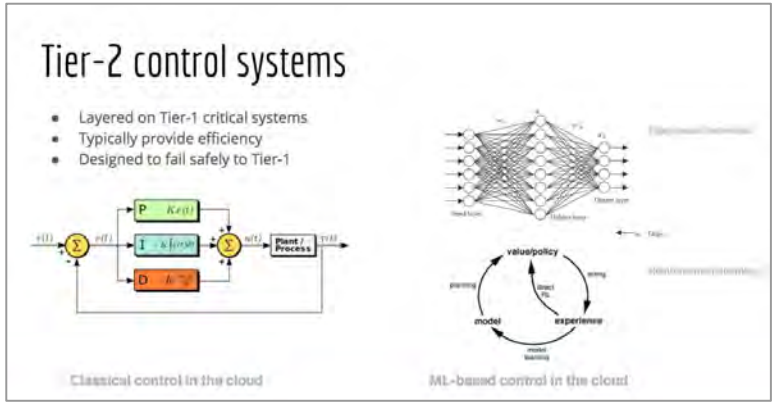
Planning for power

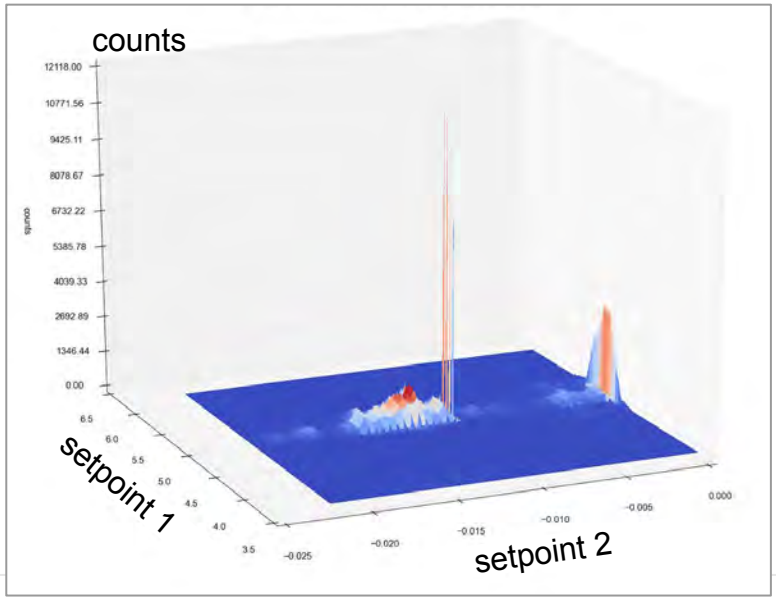
Planning for power

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.**





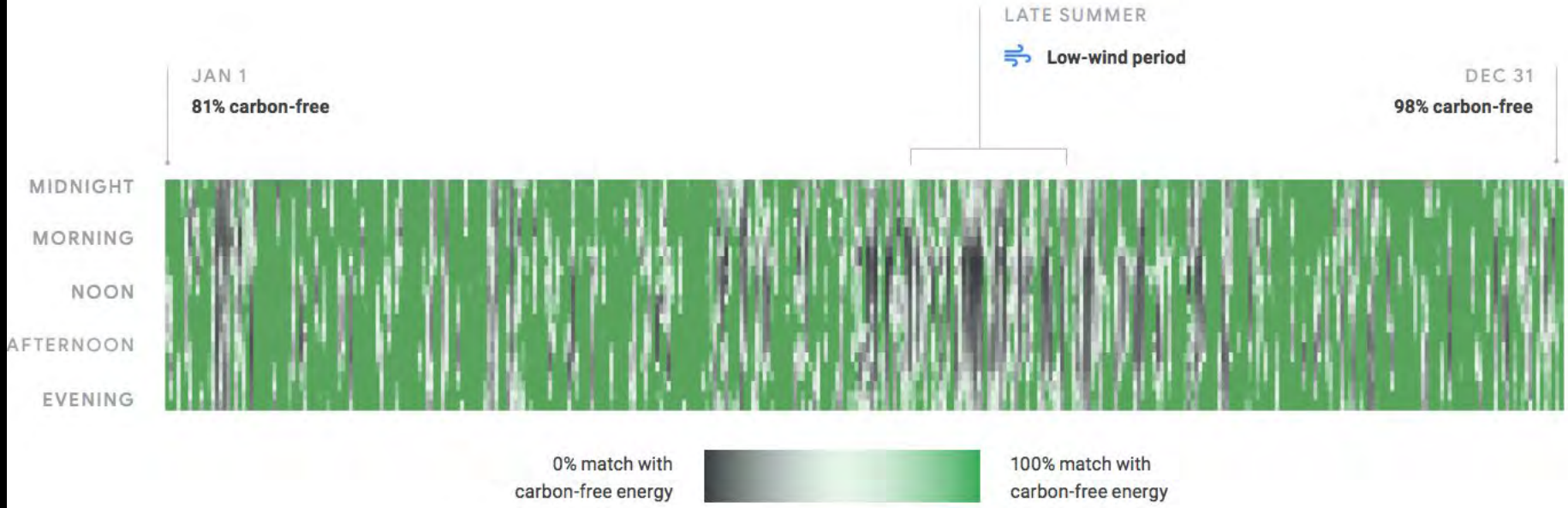
Graphics by Steve Webster

We reached 100% renewable energy balance in 2017

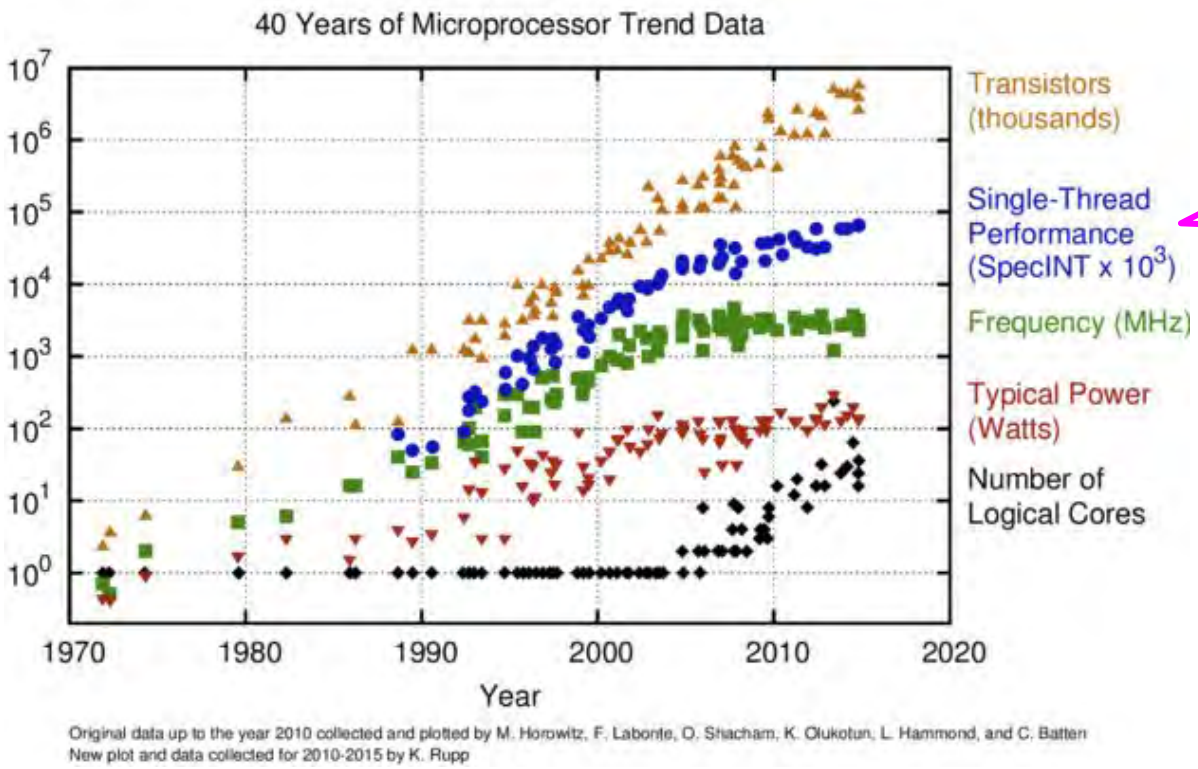
Every hour of electricity use at Iowa data center

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?



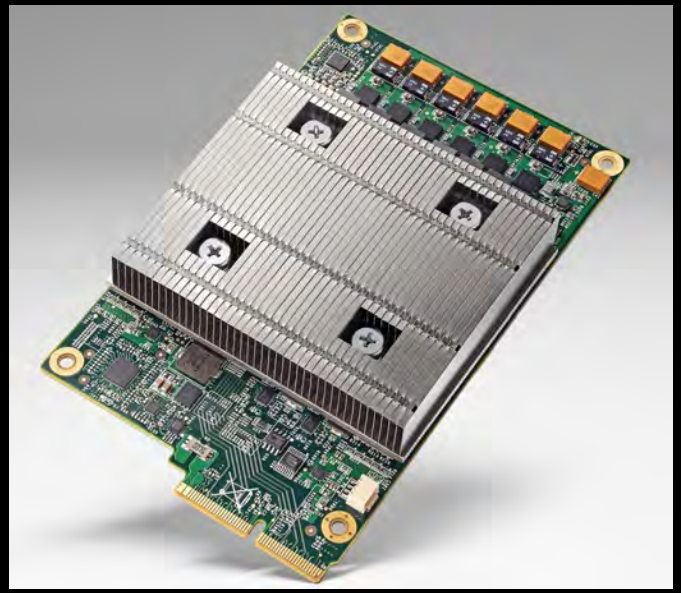
Single-core performance plateauing after decades of exponential growth

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

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

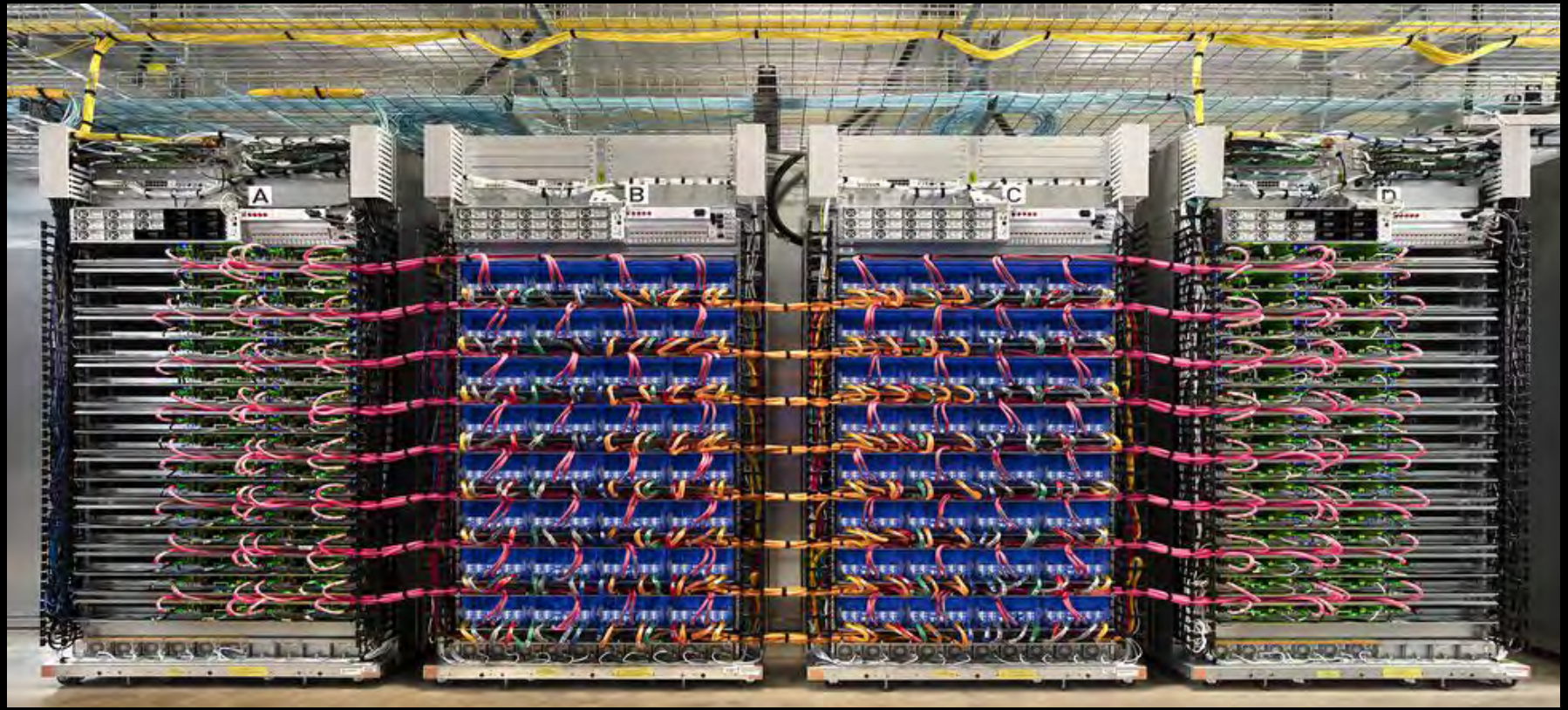
2015: Tensor Processing Unit (TPU)

2017: **Cloud TPU**
180 teraflops of computation
64 GB of HBM memory

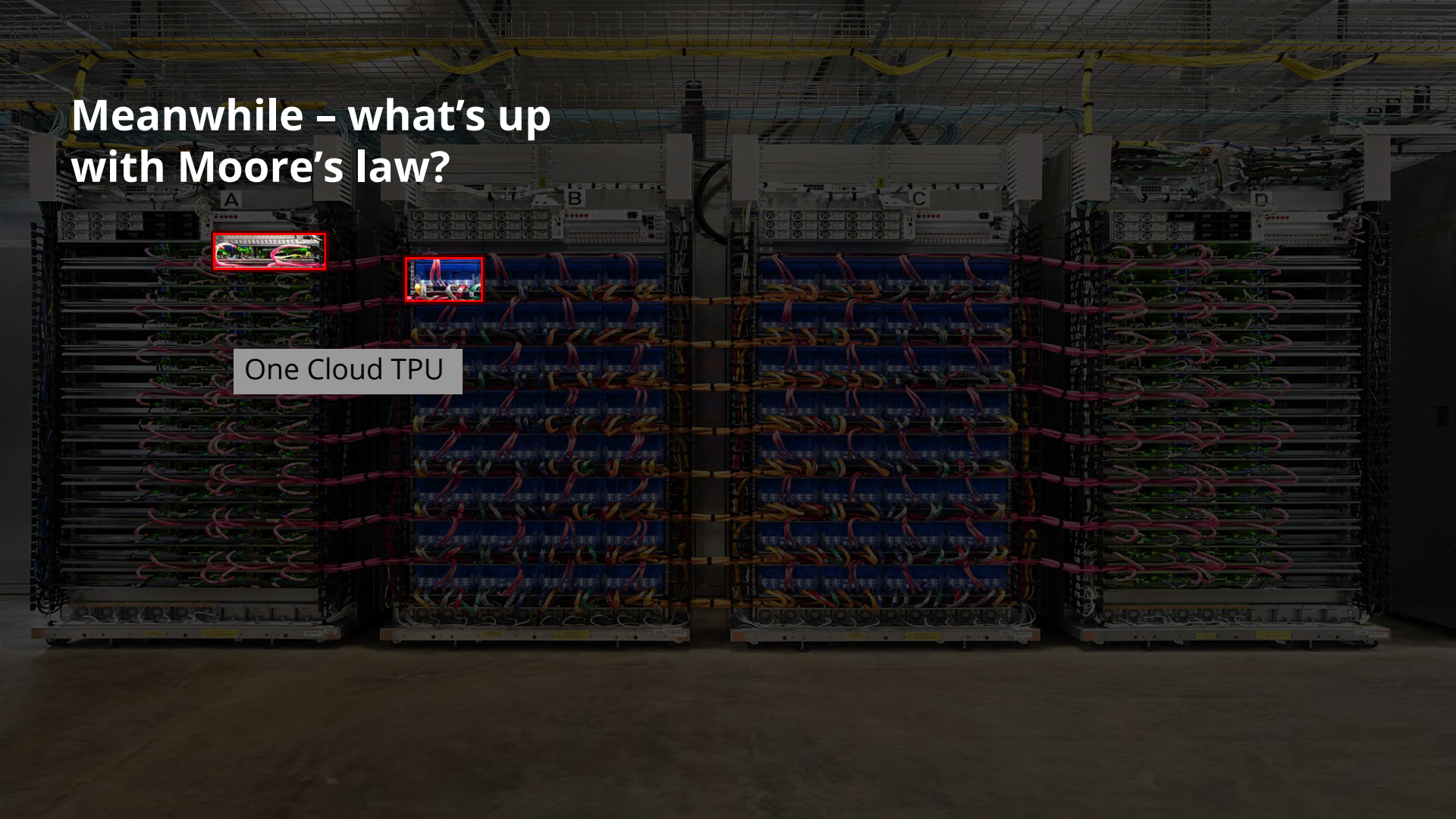


Meanwhile – what's up with Moore's law?

2017: **TPUv2**: Tensor Processing Unit **pods**
11.5 petaflops; 4 TB HBM
2-D toroidal mesh network

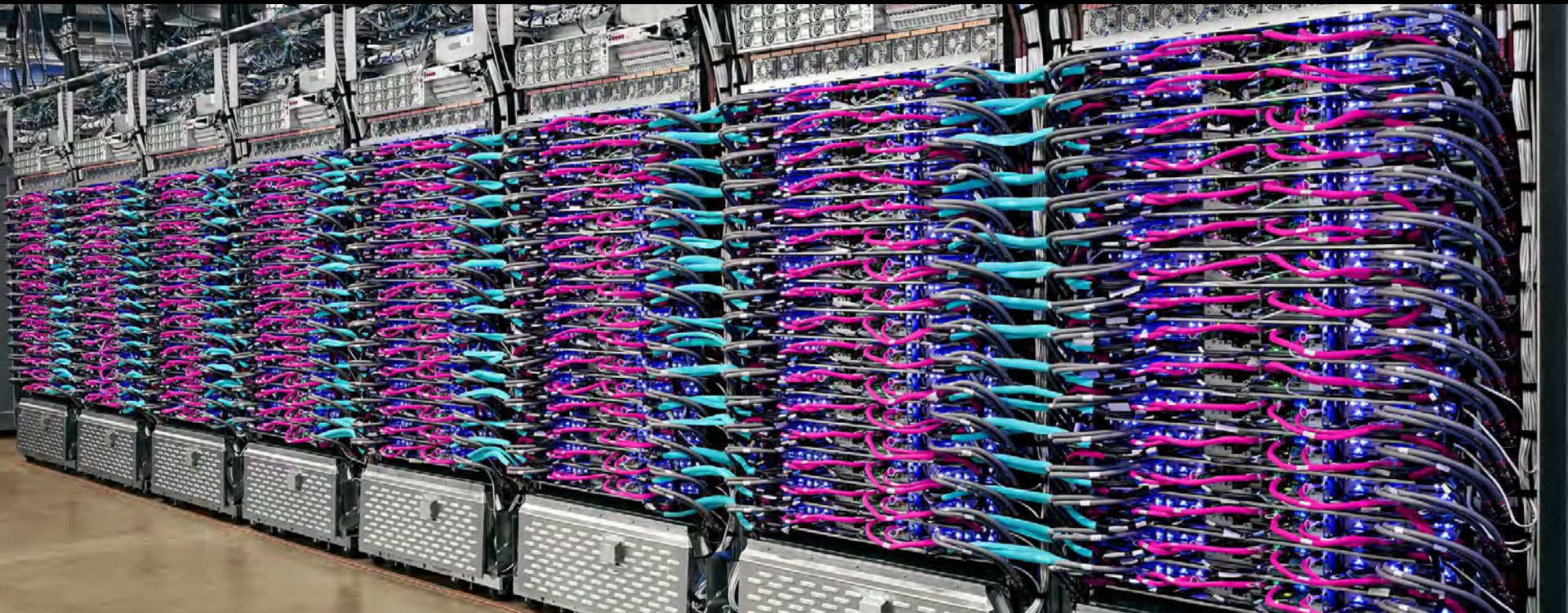


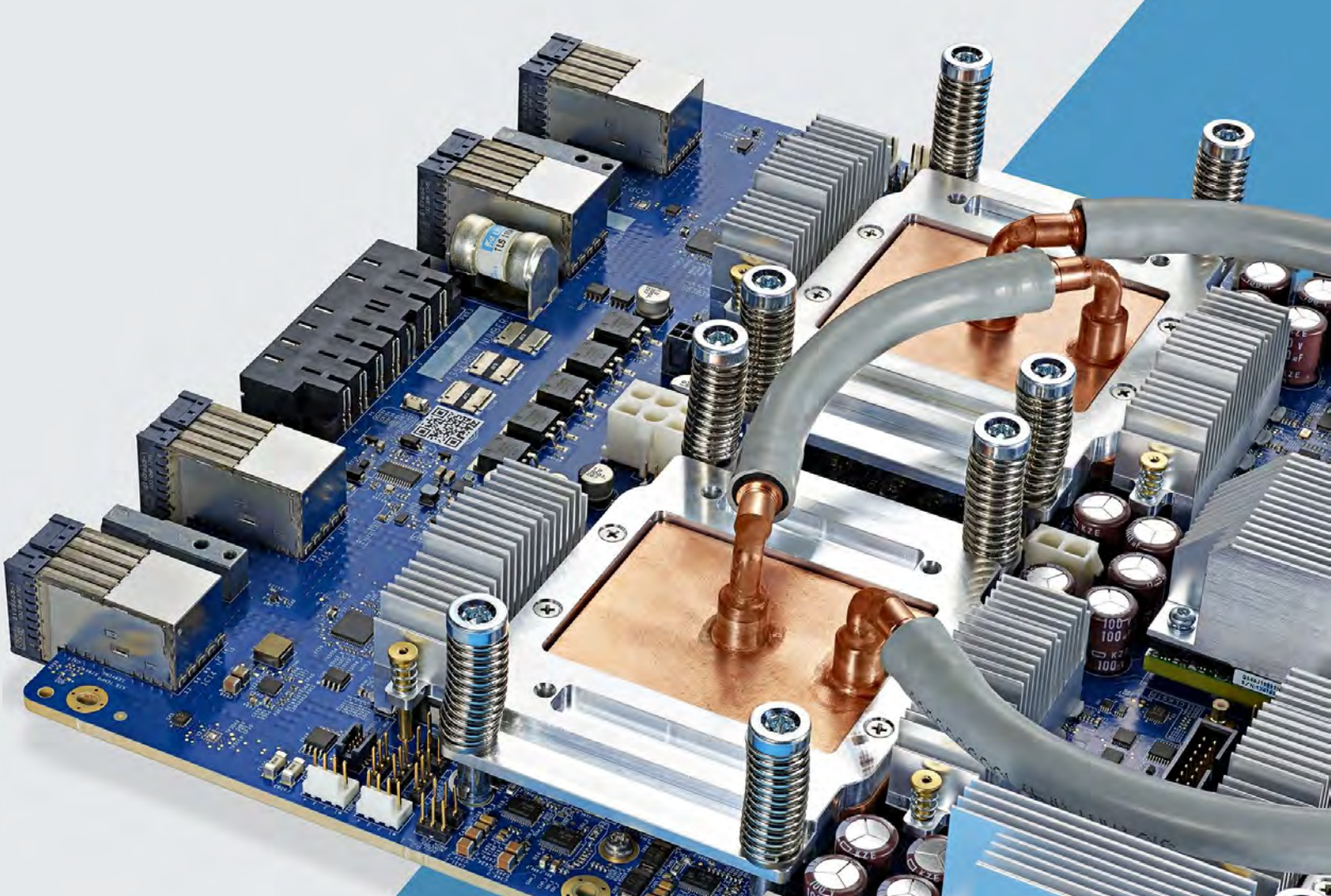
Meanwhile – what's up with Moore's law?



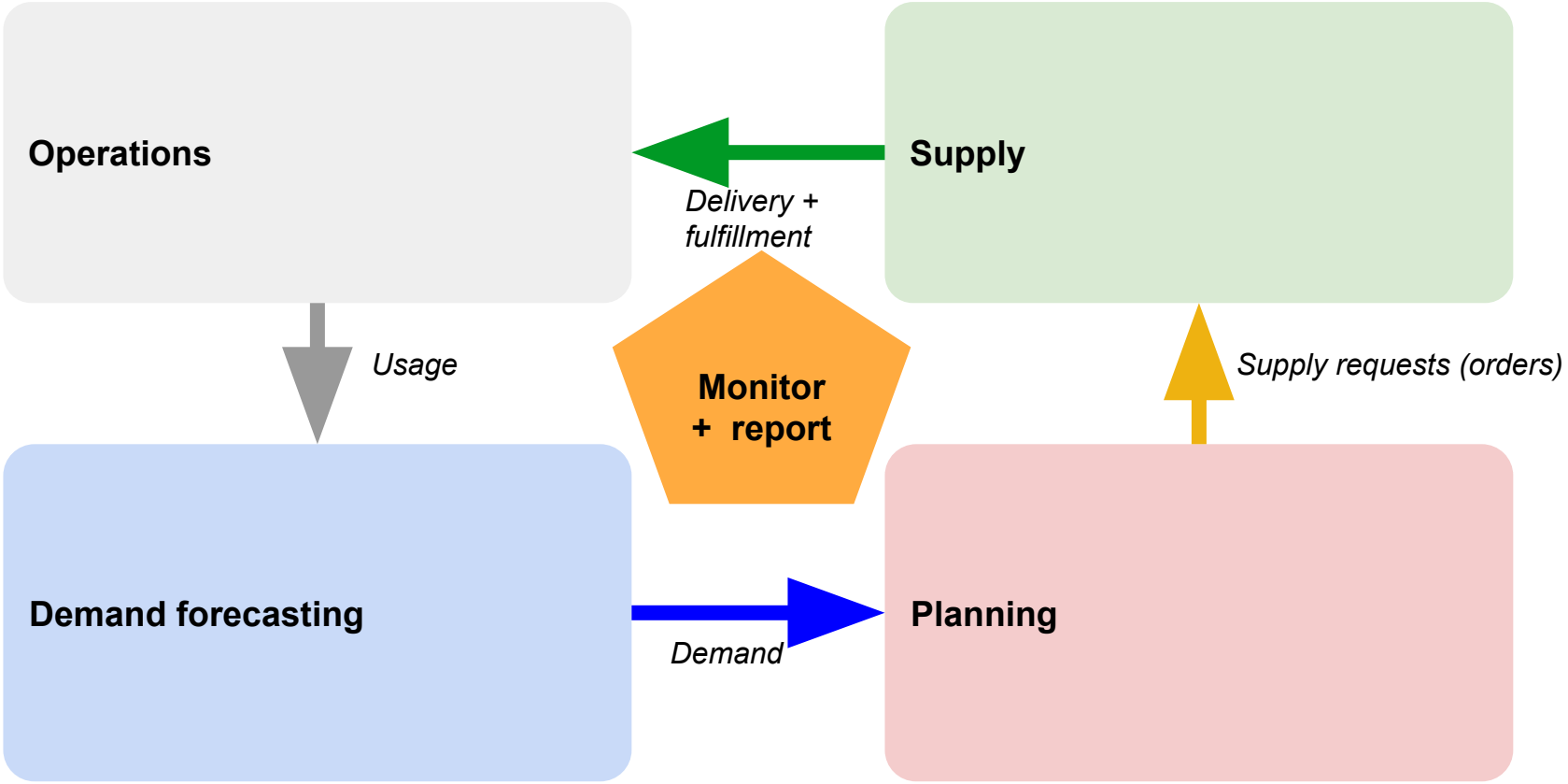
Meanwhile – what's up with Moore's law?

2018: **TPUv3**
> 100 petaflops!
> 8x the performance of a TPUv2 pod

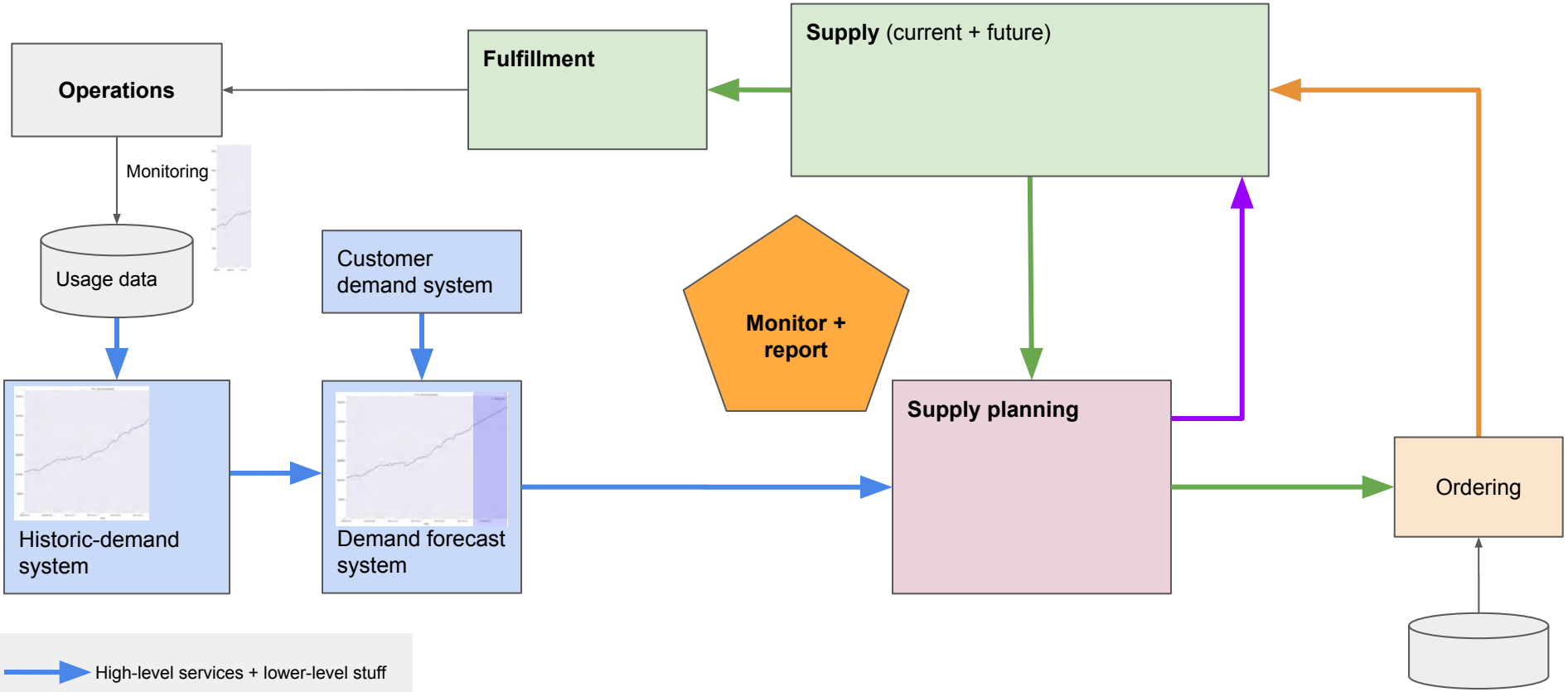




Putting it all together

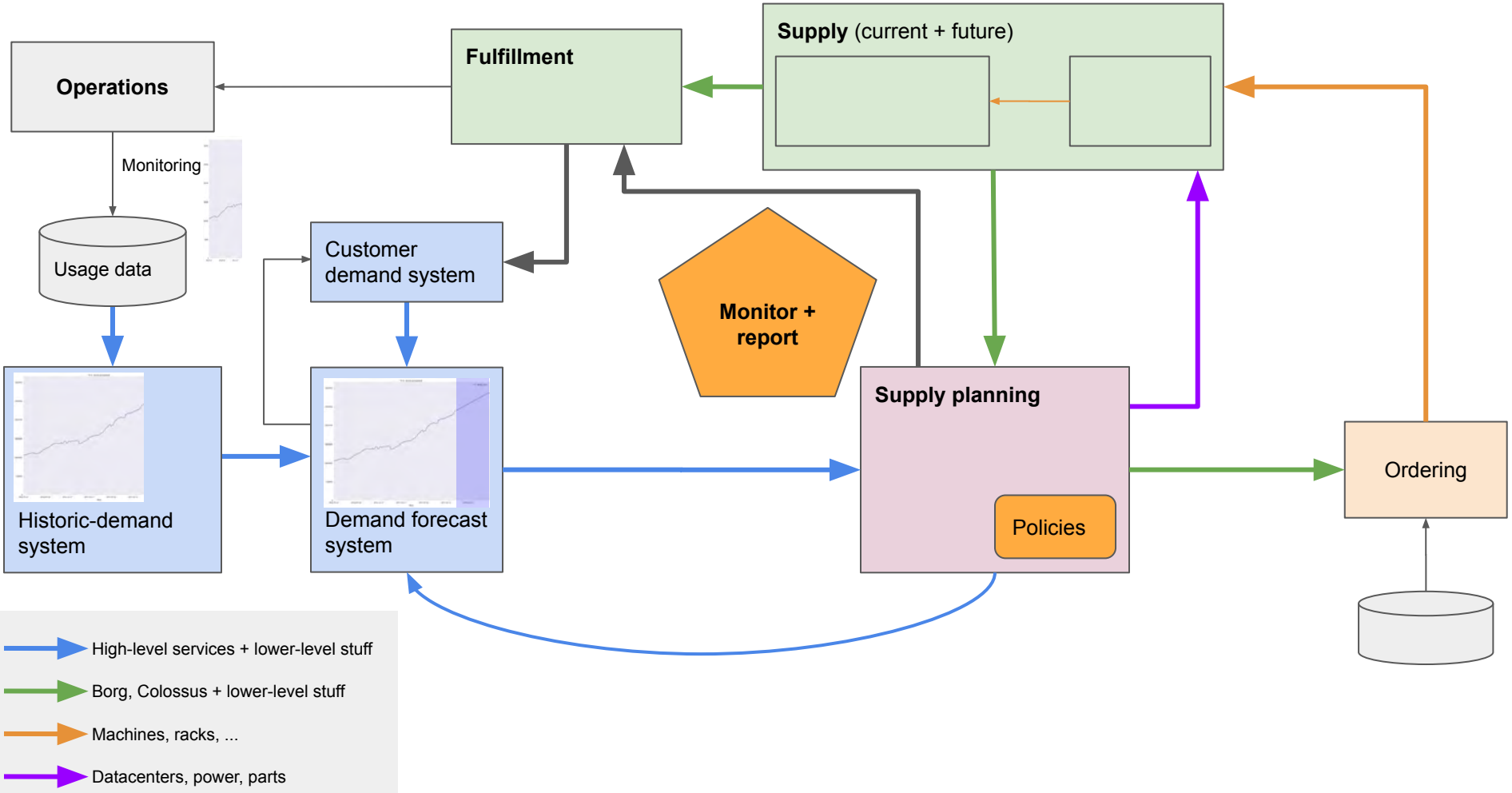


Putting it all together – a few more details

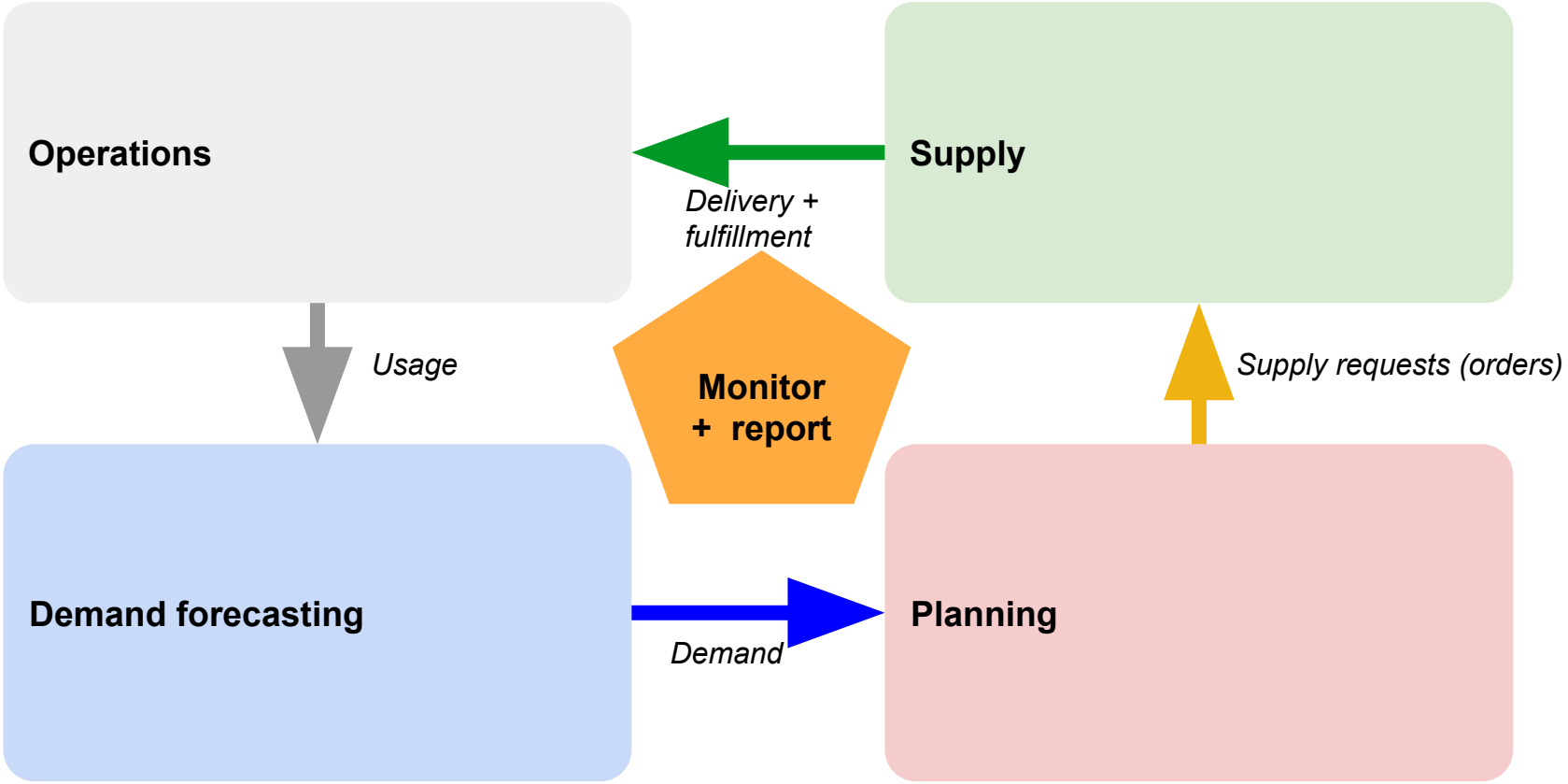


- High-level services + lower-level stuff
- Borg, Colossus + lower-level stuff
- Machines, racks, ...
- Datacenters, power, parts

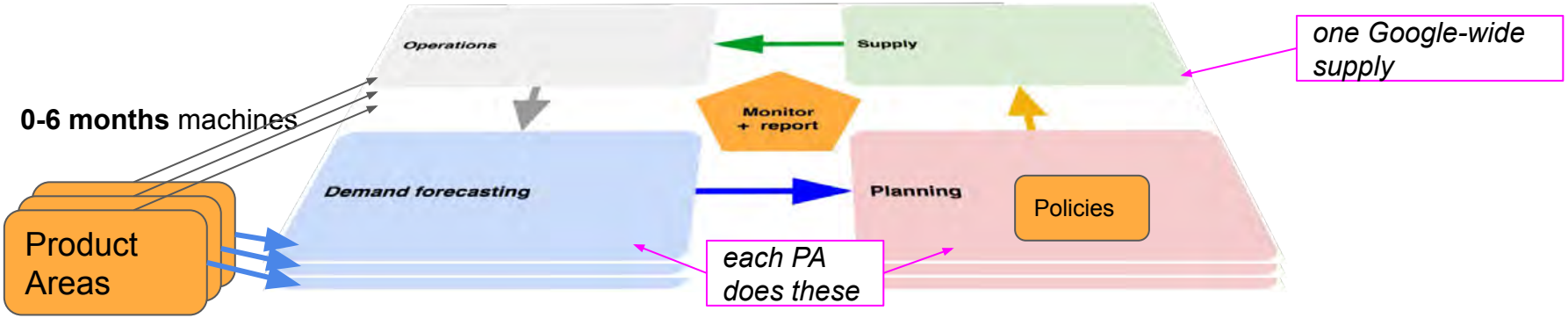
Putting it all together – a few more details



Putting it all together



Putting it all together



Putting it all together – multiple timelines

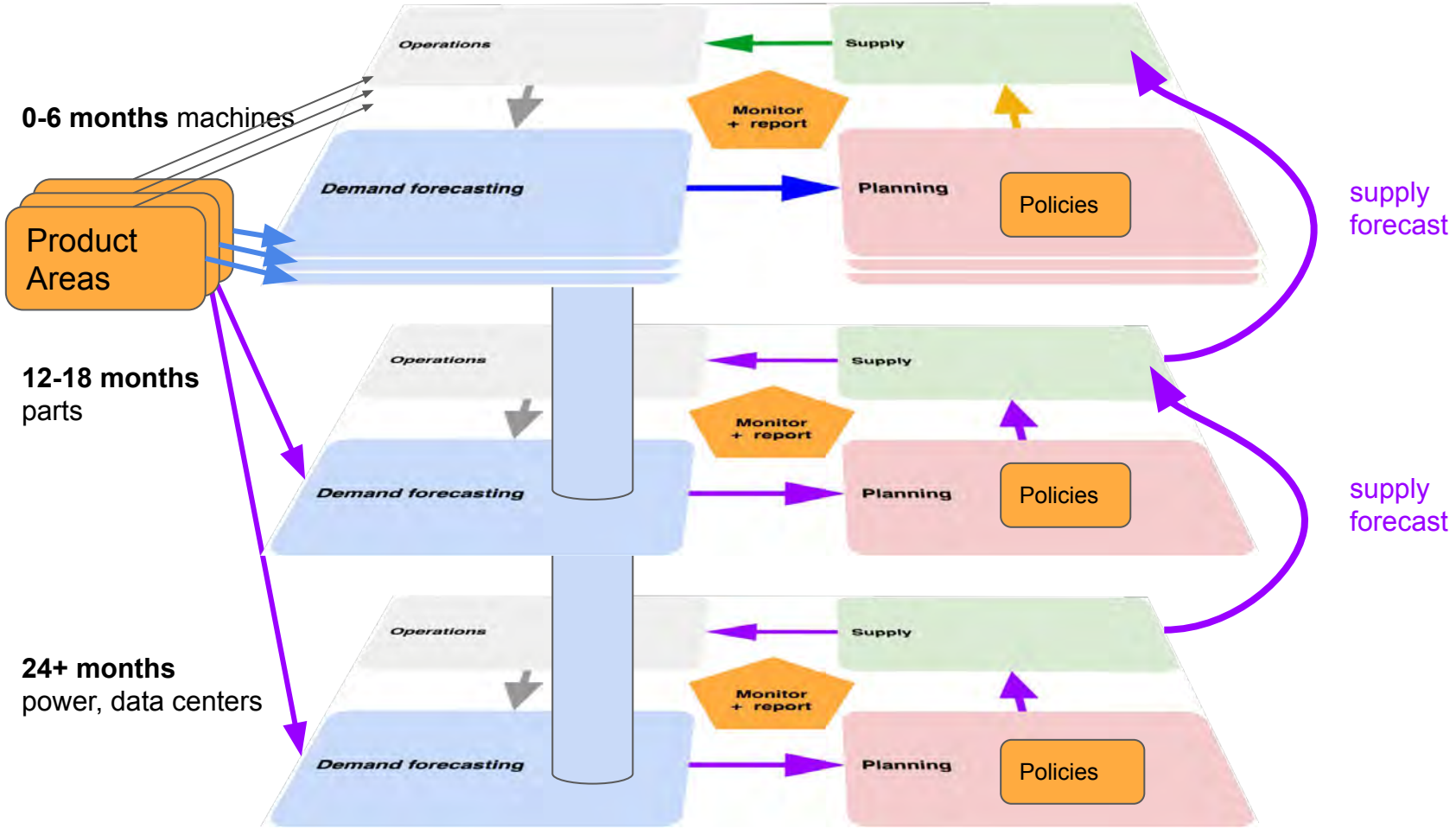


Image by Connie Zhou

2018 CapEx = $25.1B

(+$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!]

Google Cloud

johnwilkes@google.com