facebook

Facebook and Open Compute

Designing for efficiency and scale

Charlie Manese
Infrastructure Engineering
SC14 EE HPC WG Workshop - Nov. 17, 2014

Connect everyone

Understand the world

Build the knowledge economy

Driven by a need to efficiently scale

Facebook’s scale

- 1.35 Billion monthly users
- 864 Million people use Facebook each day
- 1.12 Billion Mobile monthly users

- 4.75 Billion Content Items shared daily

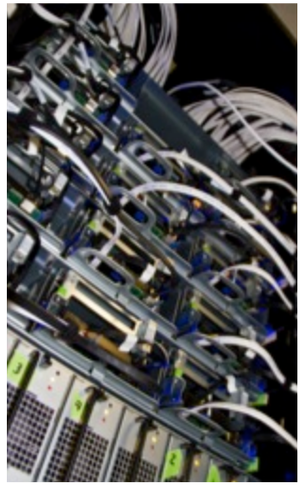
- 350 Million Photos uploaded per day
- 400 Billion photos uploaded and shared

- Instagram: 45 Million photos updated daily
- Instagram: 16 Billion photos shared

- 6 Billion Likes a day

- 200+ Billion friend connections

Complete data center ecosystem



Complete hardware ecosystem

Software

Network

Servers
& Storage

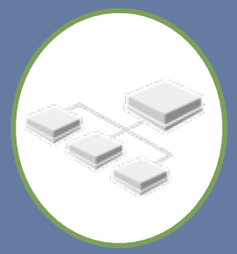
Data
Center

HipHop Virtual Machine

5x – 6x faster than Zend

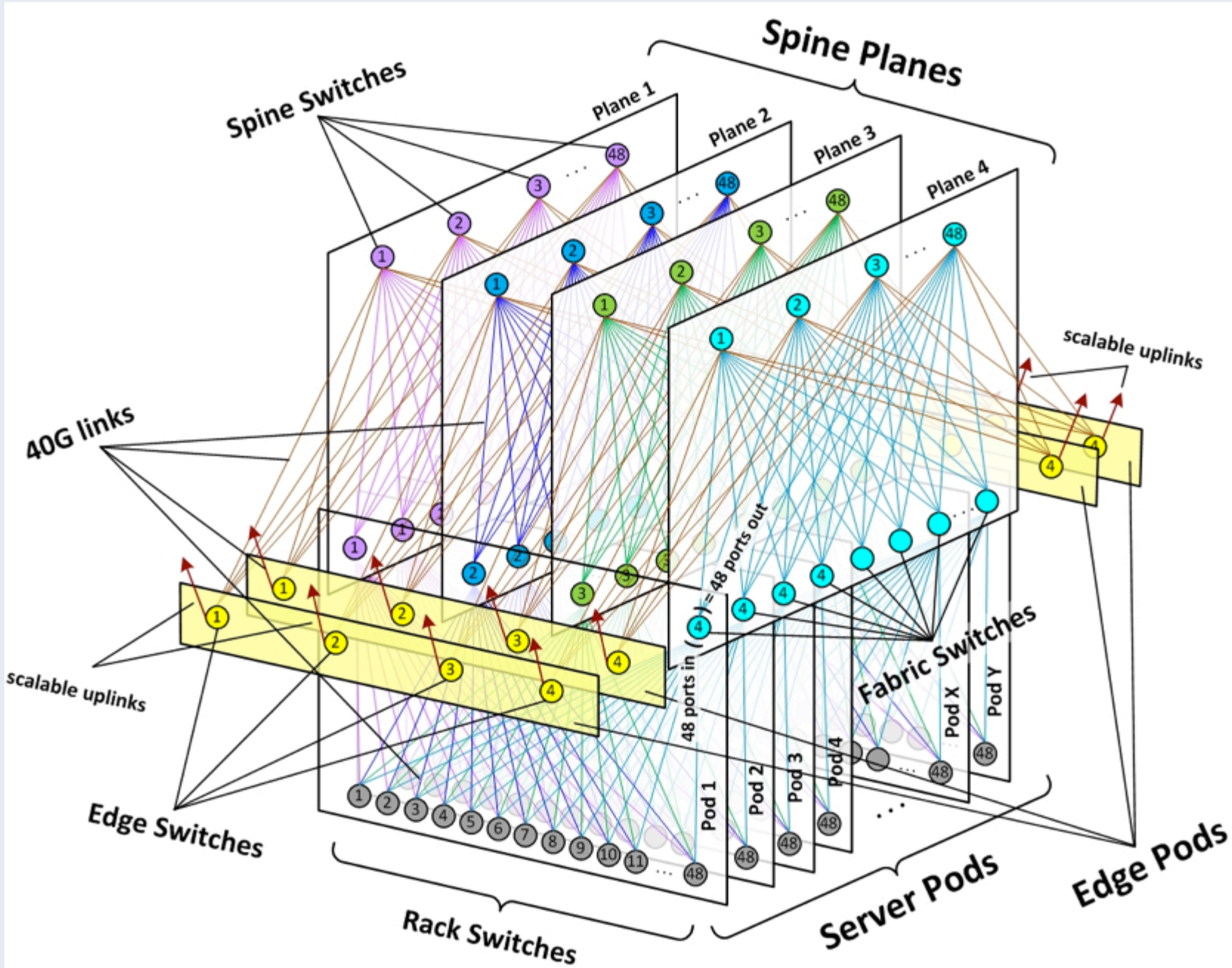
Software

Network



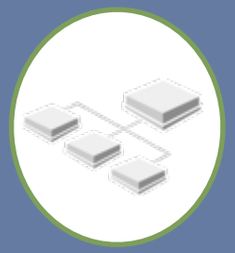
Servers
& Storage

Data
Center



Software

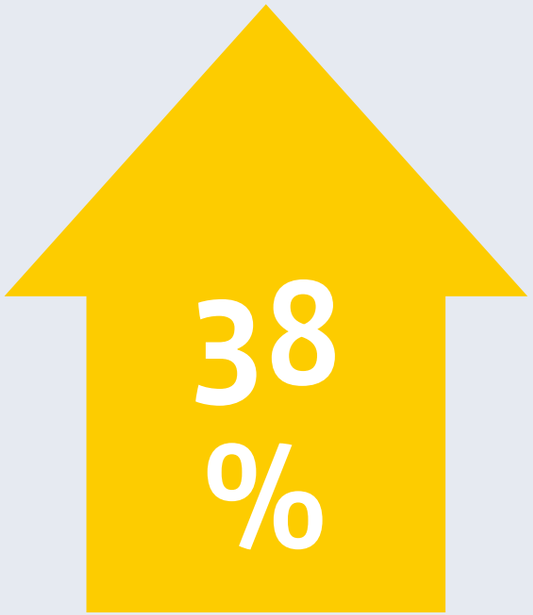
Network



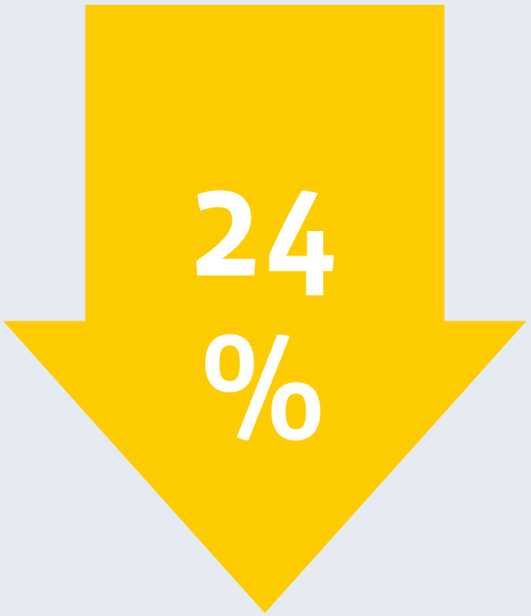
Servers & Storage

Data Center

Original OCP designs



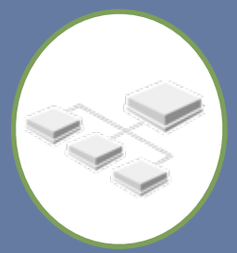
Energy Efficiency



Cost

Software

Network

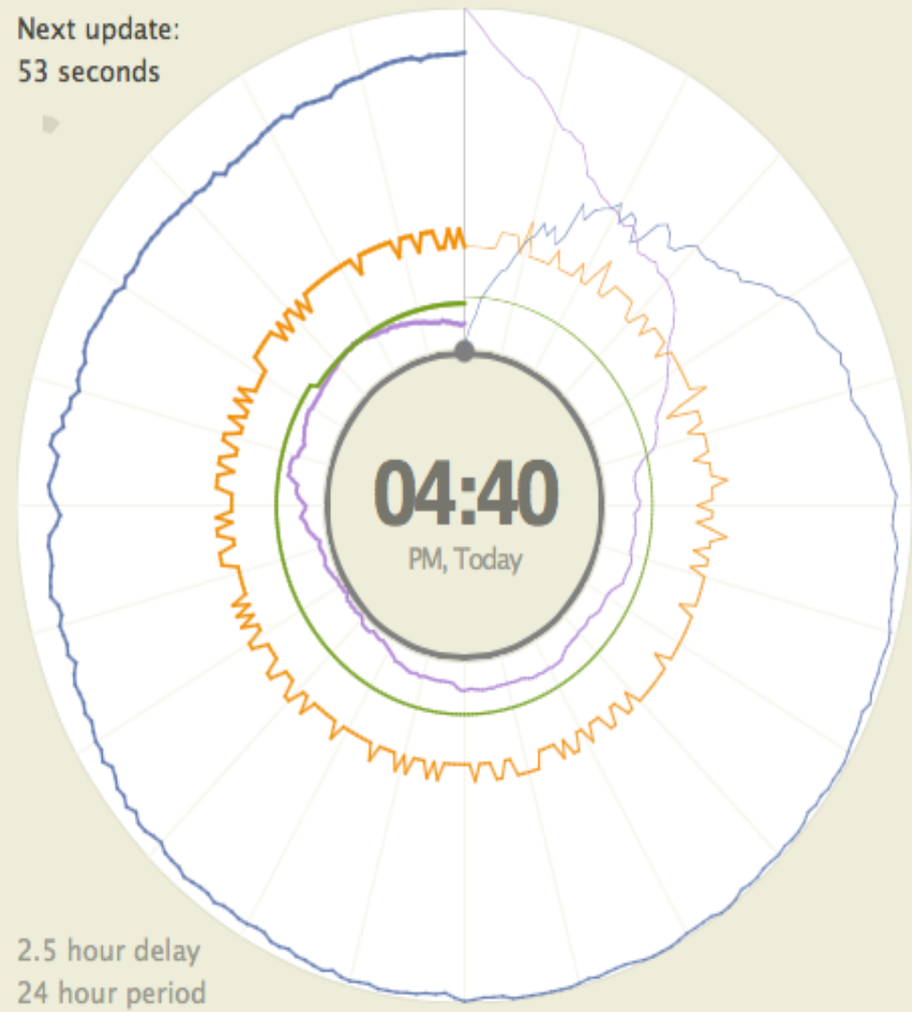


Servers & Storage

Data Center

Prineville, OR Data Center

Dashboard: PUE & WUE



Power Usage Effectiveness (PUE) **1.06**

Water Usage Effectiveness (WUE) **0.20** L/kWh

Humidity (Outdoors) **91%** /100%

Temperature (Outdoors) **28°** F /-2.2°C

Annualized Numbers — The chart above shows real-time PUE, WUE, temperature and humidity for Facebook's Prineville data center. The numbers to the right are the Prineville data center trailing 12-month PUE and WUE as of the end of September 2013.

PUE **1.08** TTM

WUE **0.49** L/kWh

$1.2 billion

DATA CENTER

Facebook Greenfield Datacenter

Goal

- Design and build the most efficient datacenter eco-system possible

Control

- Application
- Server configuration
- Datacenter design

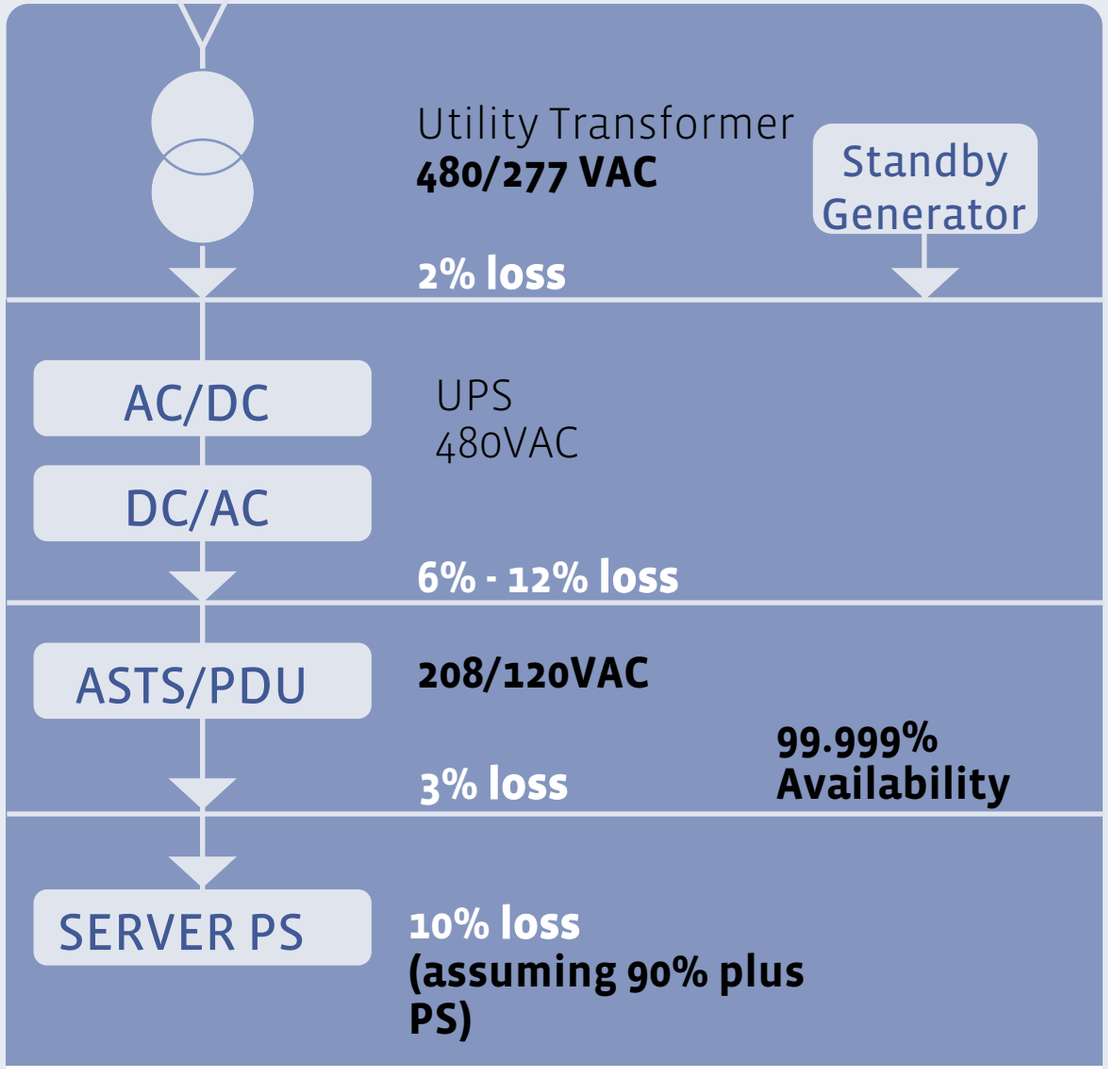
Sites

- Prineville, OR
- Forest City, NC
- Luleå, Sweden
- Altoona, IA

Electrical Overview

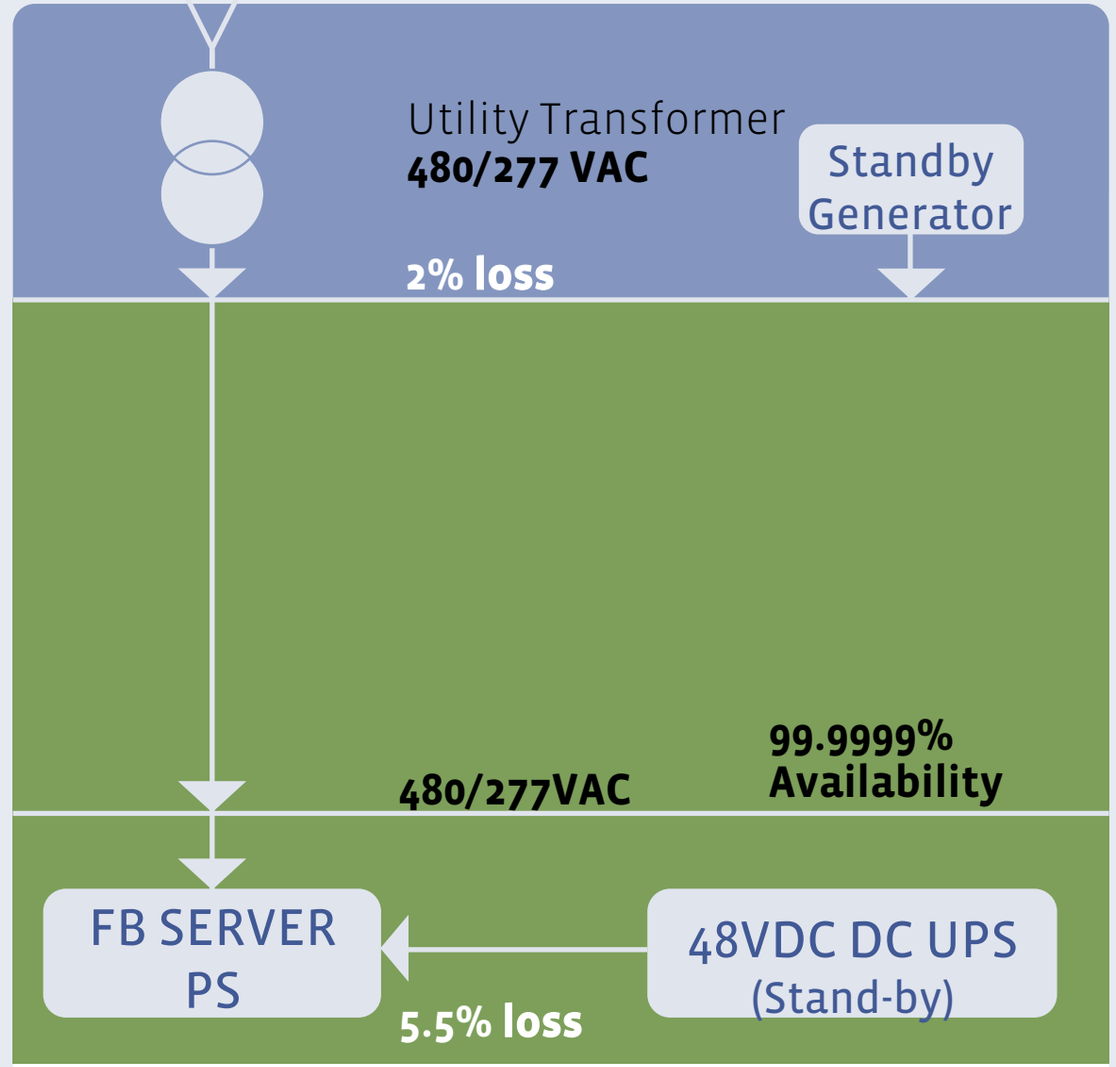
- **Eliminate** 480V to 208V transformation
  - Used 480/277VAC distribution to IT equipment
- **Remove** centralized UPS
  - Implemented 48VDC UPS System
- **Result** a highly efficient electrical system and small failure domain

Typical Power



Total loss up to server:
21% to 27%

Prineville Power

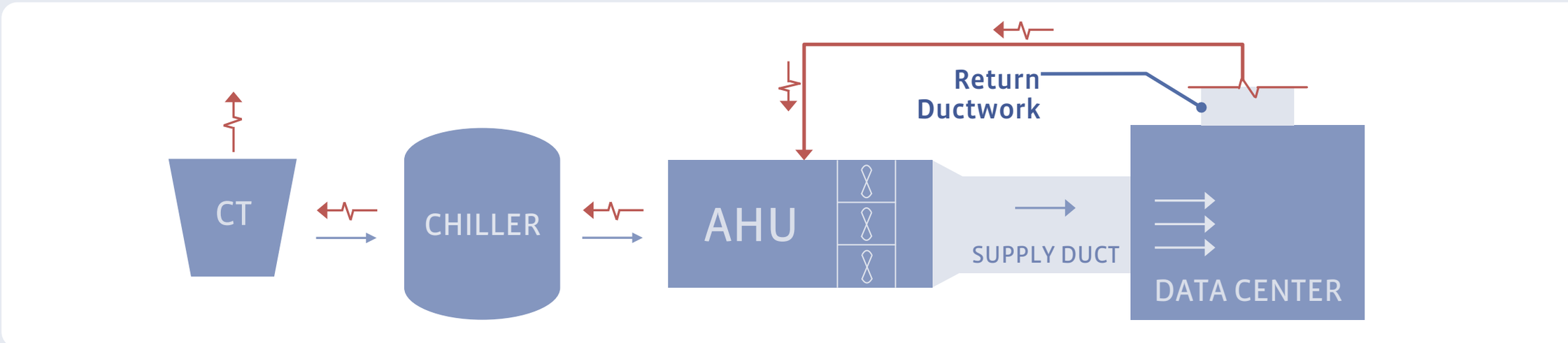


Total loss up to server:
7.5%

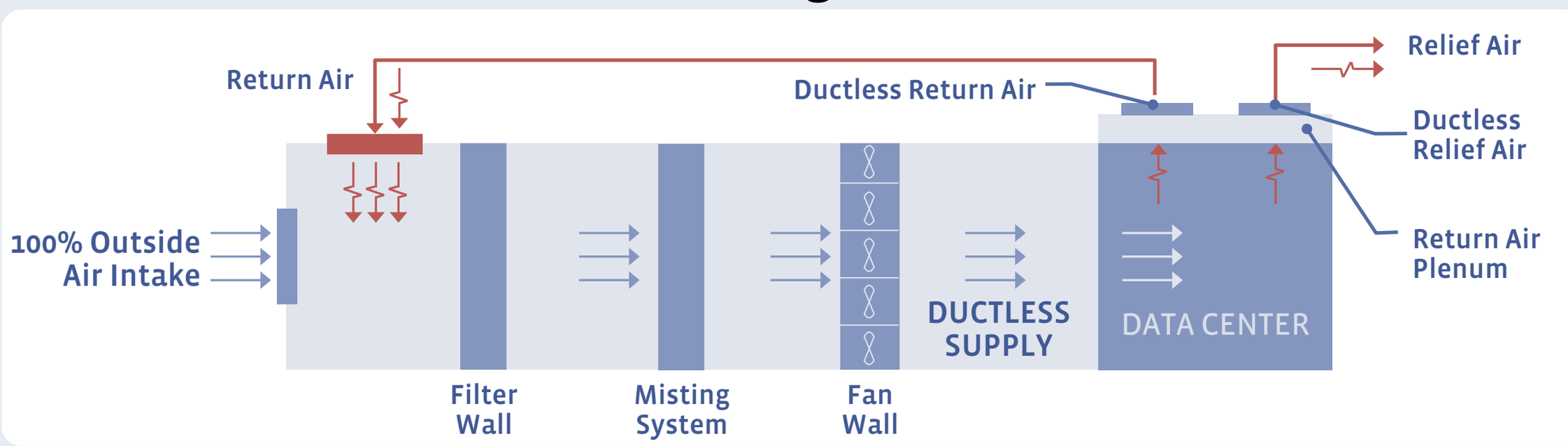
Mechanical Overview

- **Removed**
  - Centralized chiller plant
  - HVAC ductwork
- **System Basis of Design**
  - ASHRAE Weather Data: N=50 years
  - TC9.9 2008: Recommended Envelopes
- **Built-up penthouse air handling system**
- **Server waste heat is used for office space heating**

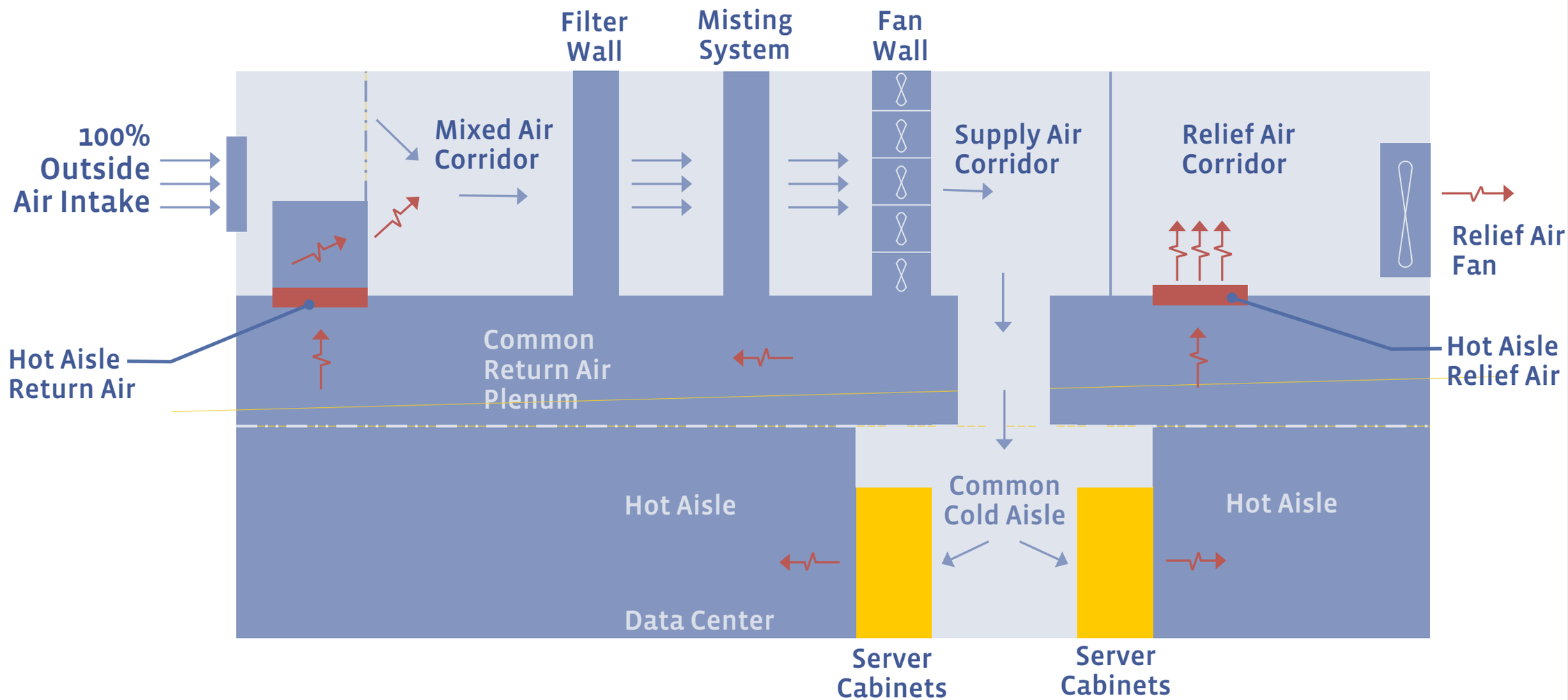
Typical Datacenter Cooling



Prineville Datacenter Cooling

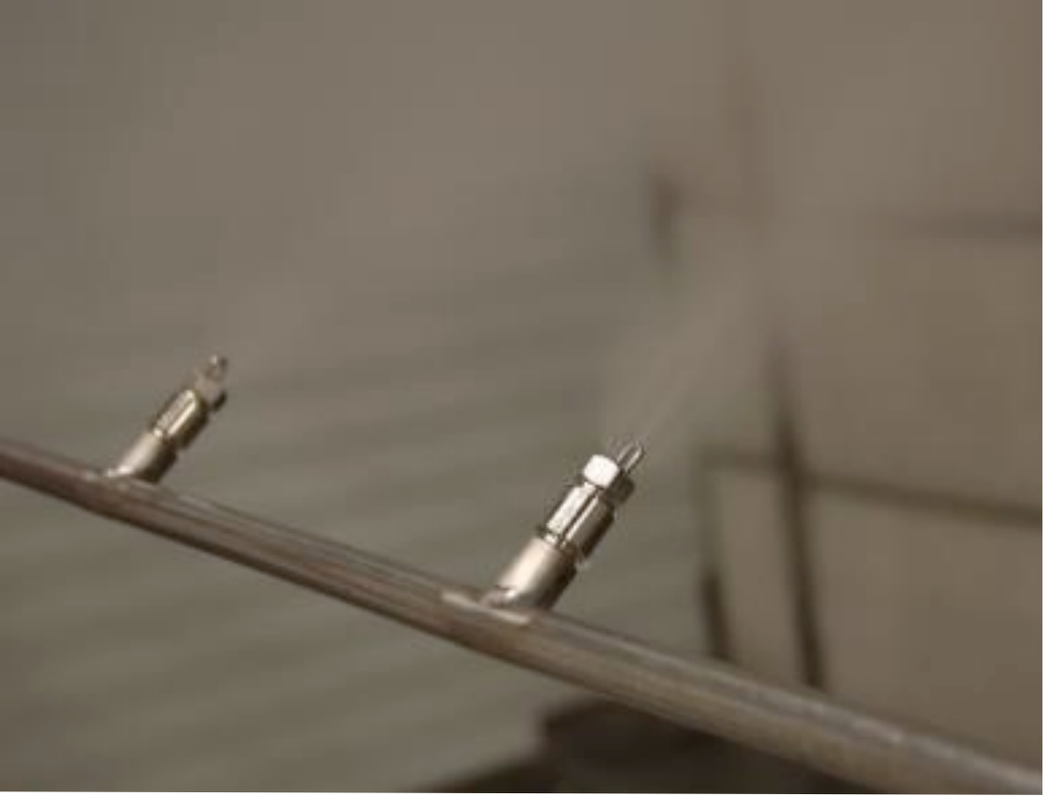


PRN Datacenter Cooling



Misting

- Reverse Osmosis water treatment required
- Central water distribution system
- 99 stages of control
- Local pump skids per AHU line up
- Mist eliminator is required

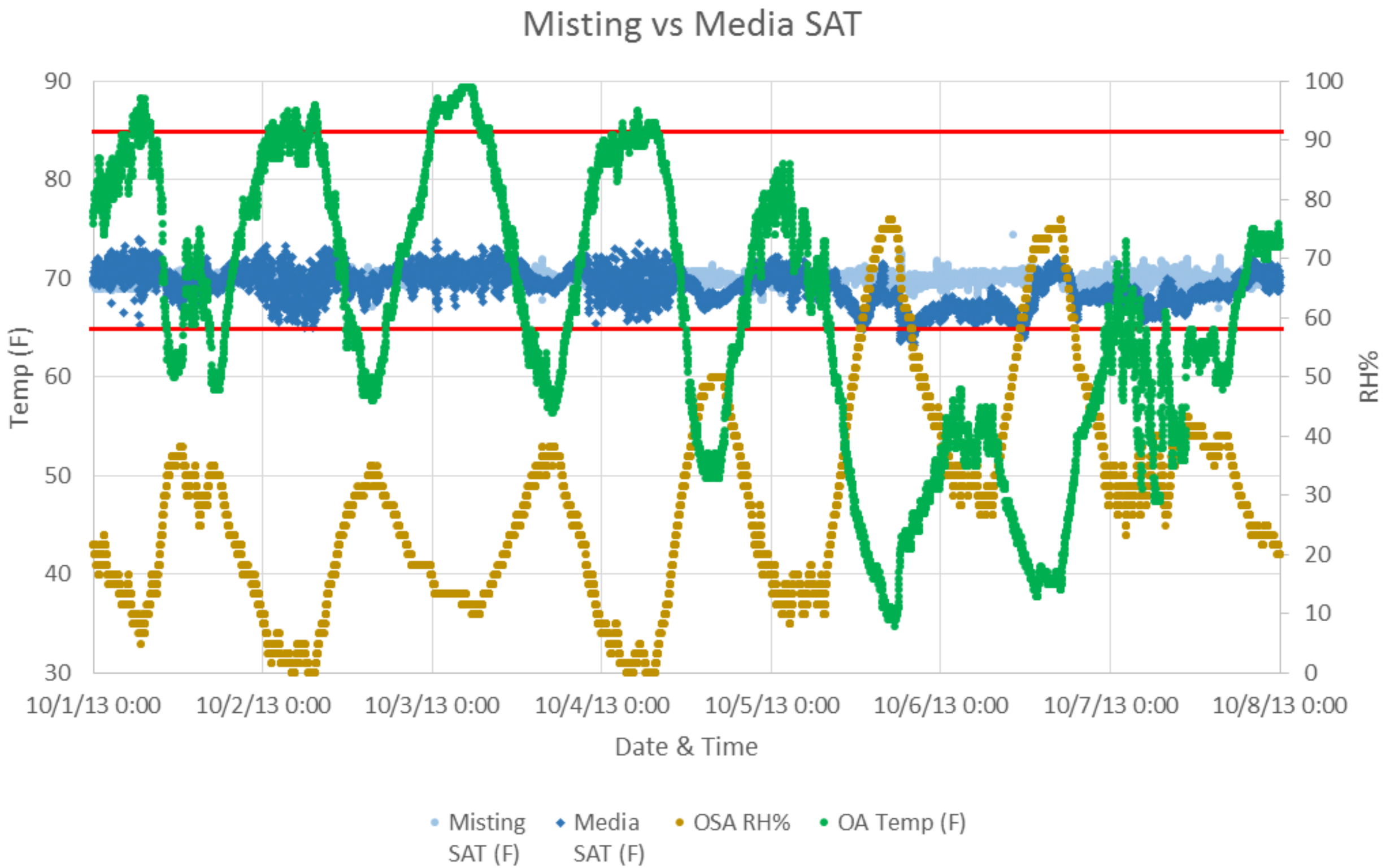


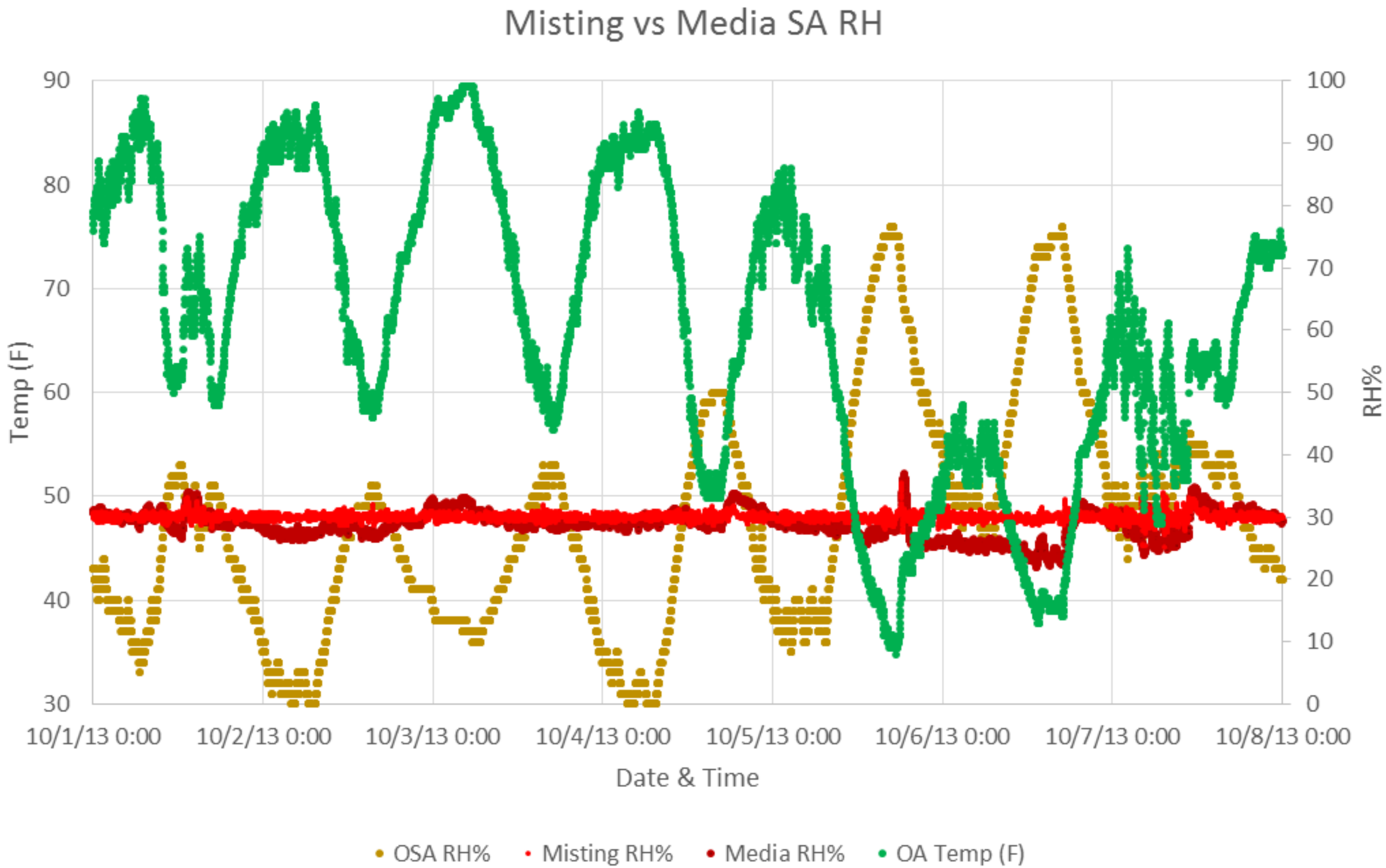
Wetted Media

- No Reverse Osmosis water treatment required
- Central water distribution system
- 4 stages of control with face & by-pass dampers
- Selectable modes of operation
  - Time Staged
  - Face/bypass
  - Hybrid

Direct Evaporative System Comparison

Description	Misting	Media	Remarks
Reverse Osmosis System	Required	Not required	
Reverse Osmosis Plant Space	Required	Not required	
Piping material	Stainless	CPVC	
Absorption Distance	6-10ft	8-12 inches	
Misting Nozzle Adjustment	Required	Not required	
Misting Nozzles Clogging	Yes	No	
Capacity Control	Good	Acceptable	
Water Consumption	Slightly higher	Slightly lower	RO, Softener, Carbon Filter Systems blow-down
System Consumable	No	Yes	Media needs replacement
Particles Removal Capability	Less favorable	More Favorable	
Pressure Drop	equal	equal	Media vs. Mist eliminator
CAPEX	High	Low	24% savings with media
OPEX	Slightly Higher	Slightly lower	RO vs. Media replacement Pump energy


Misting vs Media SAT
Temp (F)
RH%
90
80
70
60
50
40
30
100
90
80
70
60
50
40
30
20
10
0
10/1/13 0:00
10/2/13 0:00
10/3/13 0:00
10/4/13 0:00
10/5/13 0:00
10/6/13 0:00
10/7/13 0:00
10/8/13 0:00
Date & Time
Misting SAT (F)
Media SAT (F)
OSA RH%
OA Temp (F)


Misting vs Media SA RH
Temp (F)
RH%
90
80
70
60
50
40
30
100
90
80
70
60
50
40
30
20
10
0
10/1/13 0:00
10/2/13 0:00
10/3/13 0:00
10/4/13 0:00
10/5/13 0:00
10/6/13 0:00
10/7/13 0:00
10/8/13 0:00
Date & Time
OSA RH%
Misting RH%
Media RH%
OA Temp (F)

Cooling system conclusions

- Wetted media system outperforms misting systems
  - Lower CapEx by 24%
  - Lower OpEx
  - Performs within operational envelope
  - Lower water consumption by order of magnitude

Basis of Design Comparison

PRN1A1B	PRN1C1D	FRC1A1B	LLA1A1B
80°F inlet	85°F inlet	85°F inlet	85°F inlet
65% humidity	80% humidity	90% humidity	80% humidity
20°F ΔT	22°F ΔT	22°F ΔT	22°F ΔT

Modifications to Design Parameters

PRN1C1D & FRC1A1B

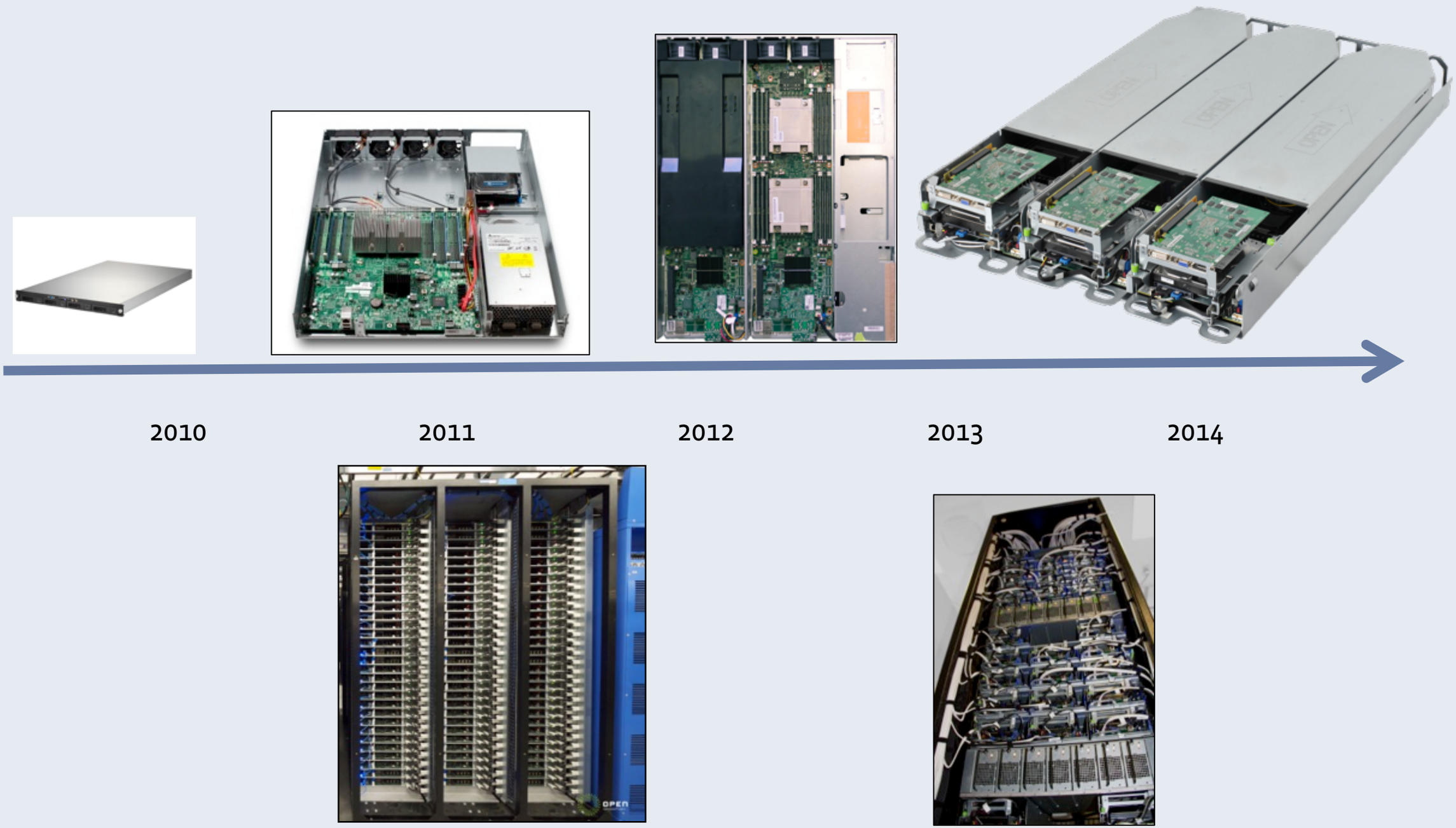
- Initial design
  - 35°F ΔT for data hall
  - Based on 100% OCP server deployment
- Final design
  - 22°F ΔT for data hall
  - Based on mix of OCP & OEM server deployment
  - Increased (8) AHU line ups to (13) to meet 22°F ΔT

RACK, SERVERS, AND STORAGE

Standard servers

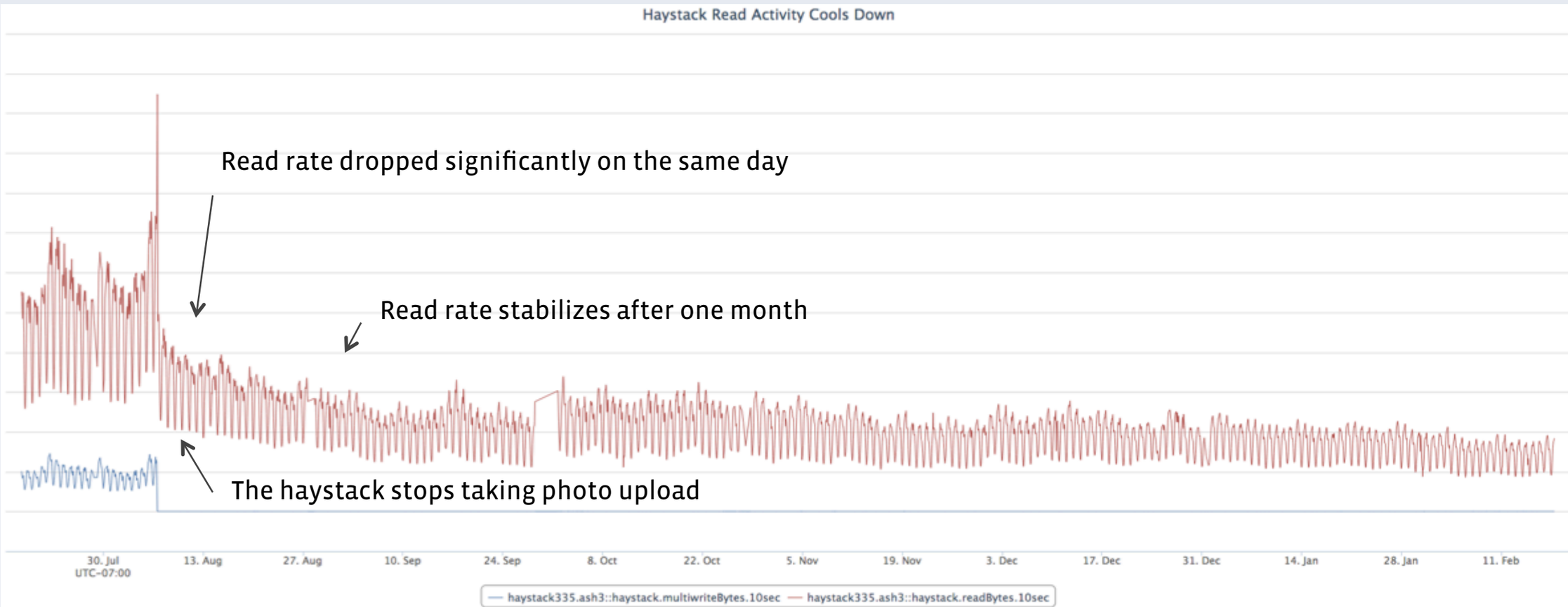
Standard Systems	I Web	III Database	IV Hadoop	V Haystack	VI Feed
CPU	2 x E5-2680v2	2 x E5-2680v2	2 x E5-2680v2	1 x E5-2660v2	2 x E5-2660v2
Memory	32GB	144GB	64GB	96GB	144GB
Disk	Boot	High IOPS 3.2 TB Flash	15 x 4TB	30 x 4TB	2TB SATA
Services	Web, Chat	Database	Hadoop	Photos, Video	Multifeed, Search, Ads

Server and rack evolution



Cold photos

Photos cool down very rapidly



What is cold storage?

Data center, storage and software for archival storage

- 8x storage density at the server level
- Data center with different power footprint
- Customized rack to accommodate power and weight
- 3 racks sharing one TOR switch
- 1 hard drive spinning on each tray at a time
- Software coping with all of the above

Six standard servers

Standard Systems	I Web	III Database	IV Hadoop	V Haystack	VI Feed	VII Cold Store
CPU	2 x E5-2680v2	2 x E5-2680v2	2 x E5-2680v2	1 x E5-2660v2	2 x E5-2660v2	2x E5-2660v2
Memory	32GB	144GB	64GB	96GB	144GB	144GB
Disk	Boot	High IOPS 3.2 TB Flash	High 15 x 4TB	High 30 x 4TB	Medium 2TB SATA	High 240 x 4TB SATA
Services	Web, Chat	Database	Hadoop	Photos, Video	Multifeed, Search, Ads	Cold Storage

Rack layout

Rack A

Position:		1:240 HDD		
41	1U	Empty		
40	1U	Empty		
39 38	2U	Empty	Winterfell	Empty
37 36	2U	Cold Storage		
35 34	2U	Cold Storage		
33 32	2U	Cold Storage		
31 30	2U	Cold Storage		
29 28	2U	Cold Storage		
27 26	2U	Cold Storage		
25 24	2U	Cold Storage		
23 22	2U	Cold Storage		
21 20 19	3U	Power Shelf		
18 17	2U	Empty	Winterfell	Empty
16 15	2U	Cold Storage		
14 13	2U	Cold Storage		
12 11	2U	Cold Storage		
10 9	2U	Cold Storage		
8 7	2U	Cold Storage		
6 5	2U	Cold Storage		
4 3	2U	Cold Storage		
2 1	2U	Cold Storage		
	WF Sub Posit	1	2	3

Rack B

Position:		1:240 HDD		
41	1U	Cisco 3064 10G Switch		
40	1U	Empty		
39 38	2U	Empty	Winterfell	Empty
37 36	2U	Cold Storage		
35 34	2U	Cold Storage		
33 32	2U	Cold Storage		
31 30	2U	Cold Storage		
29 28	2U	Cold Storage		
27 26	2U	Cold Storage		
25 24	2U	Cold Storage		
23 22	2U	Cold Storage		
21 20 19	3U	Power Shelf		
18 17	2U	Empty	Winterfell	Empty
16 15	2U	Cold Storage		
14 13	2U	Cold Storage		
12 11	2U	Cold Storage		
10 9	2U	Cold Storage		
8 7	2U	Cold Storage		
6 5	2U	Cold Storage		
4 3	2U	Cold Storage		
2 1	2U	Cold Storage		
	WF Sub Posit	1	2	3

Rack C

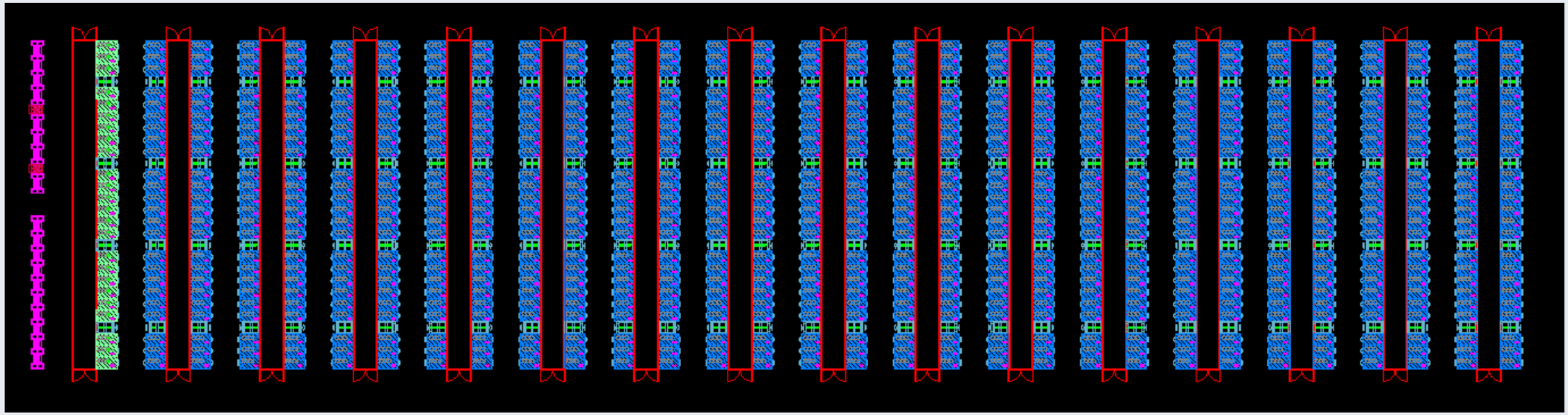
Position:		1:240 HDD		
41	1U	Empty		
40	1U	Empty		
39 38	2U	Empty	Winterfell	Empty
37 36	2U	Cold Storage		
35 34	2U	Cold Storage		
33 32	2U	Cold Storage		
31 30	2U	Cold Storage		
29 28	2U	Cold Storage		
27 26	2U	Cold Storage		
25 24	2U	Cold Storage		
23 22	2U	Cold Storage		
21 20 19	3U	Power Shelf		
18 17	2U	Empty	Winterfell	Empty
16 15	2U	Cold Storage		
14 13	2U	Cold Storage		
12 11	2U	Cold Storage		
10 9	2U	Cold Storage		
8 7	2U	Cold Storage		
6 5	2U	Cold Storage		
4 3	2U	Cold Storage		
2 1	2U	Cold Storage		
	WF Sub Posit	1	2	3

A new power density

	Watts
Storage Unit (OpenVault with 2 spinning HDD's)	70
Compute Node	300
Network Switch	200
Rack without network switch (70 x 16) + (300 x 2)	1,720
Rack with network switch (70 x 16) + (300 x 2) + 200	1,920
3 rack set (2 racks without switch + 1 rack with switch)	5,360

Cold storage data center

Data center layout



- 744 racks
- 31 rows for Cold Storage
  - 30 rows of Cold Storage system
  - 1 row protected power for cache/service racks
- 24 racks per row

Weighty issue

	kg
Storage Unit (Open Vault) with 4 TB HDD's	50
Compute Node	23
Power shelf	27
Switch and cables	20
Open Rack	230
Total (50 x 16) + (23 x 2) + 27 + 20 + 230	1,123

OpenRack for cold storage

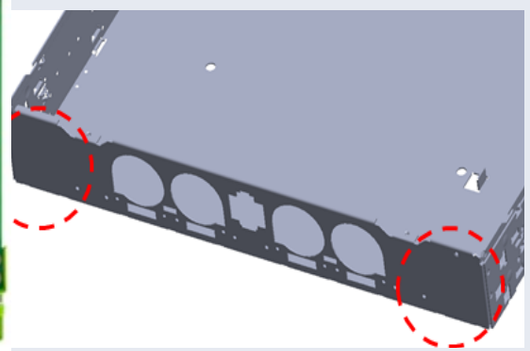
Optimized for power and weight

- Single power zone and power bar
  - Power shelf with 5 PSU modules to supply 2.8kW maximum with 4+1 redundancy
  - Rear support mechanisms on both sides of the rack

OpenVault for cold storage

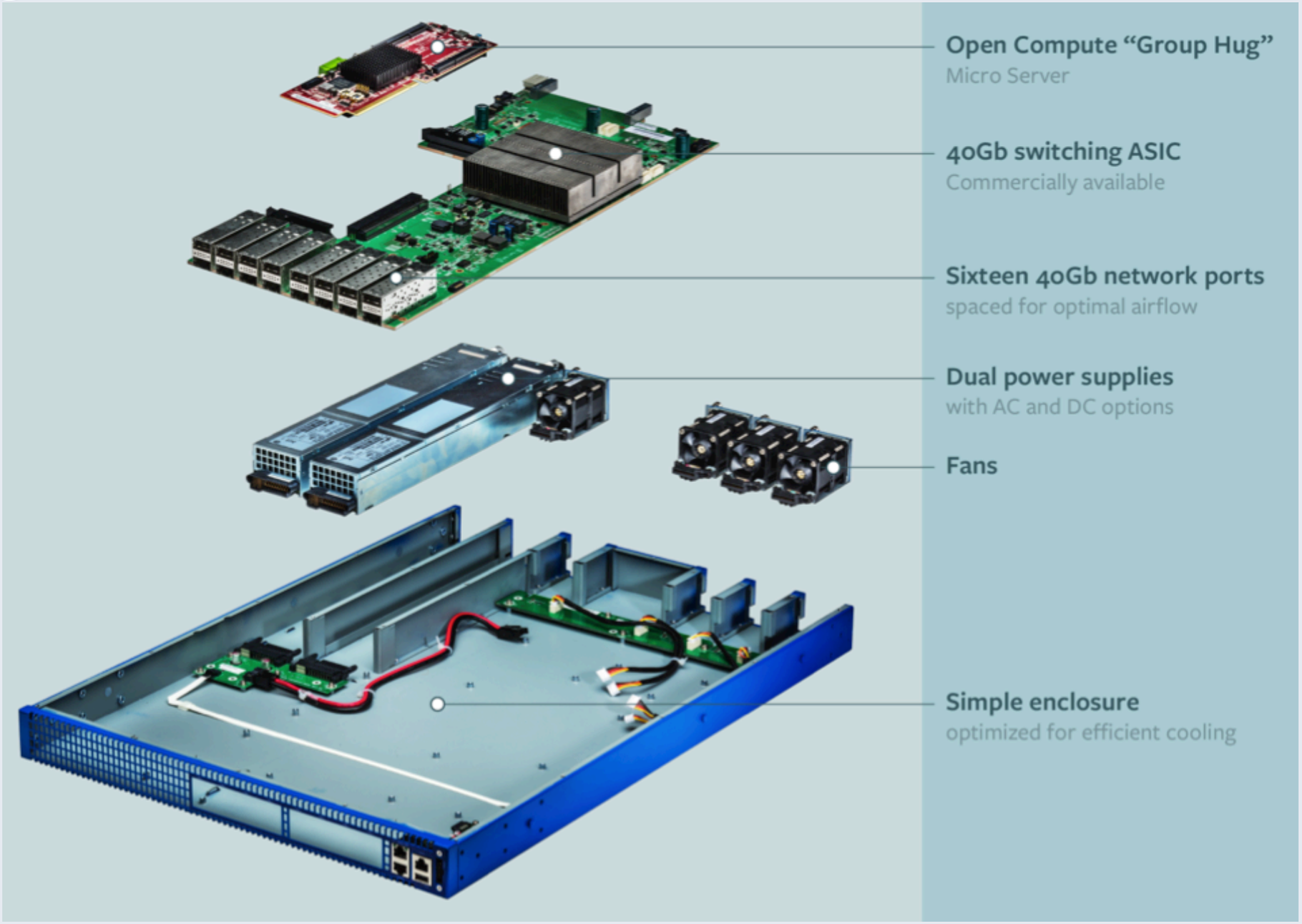
Modified to accommodate:

- Drives – only one drive spinning on a tray at a time
- Weight – removed two fans to couple with OpenRack weight support
- Thermals – lower fan pwm
- Storage architecture – Ports removed to support fan-out design
- Cables – x1 SAS fan-out cable

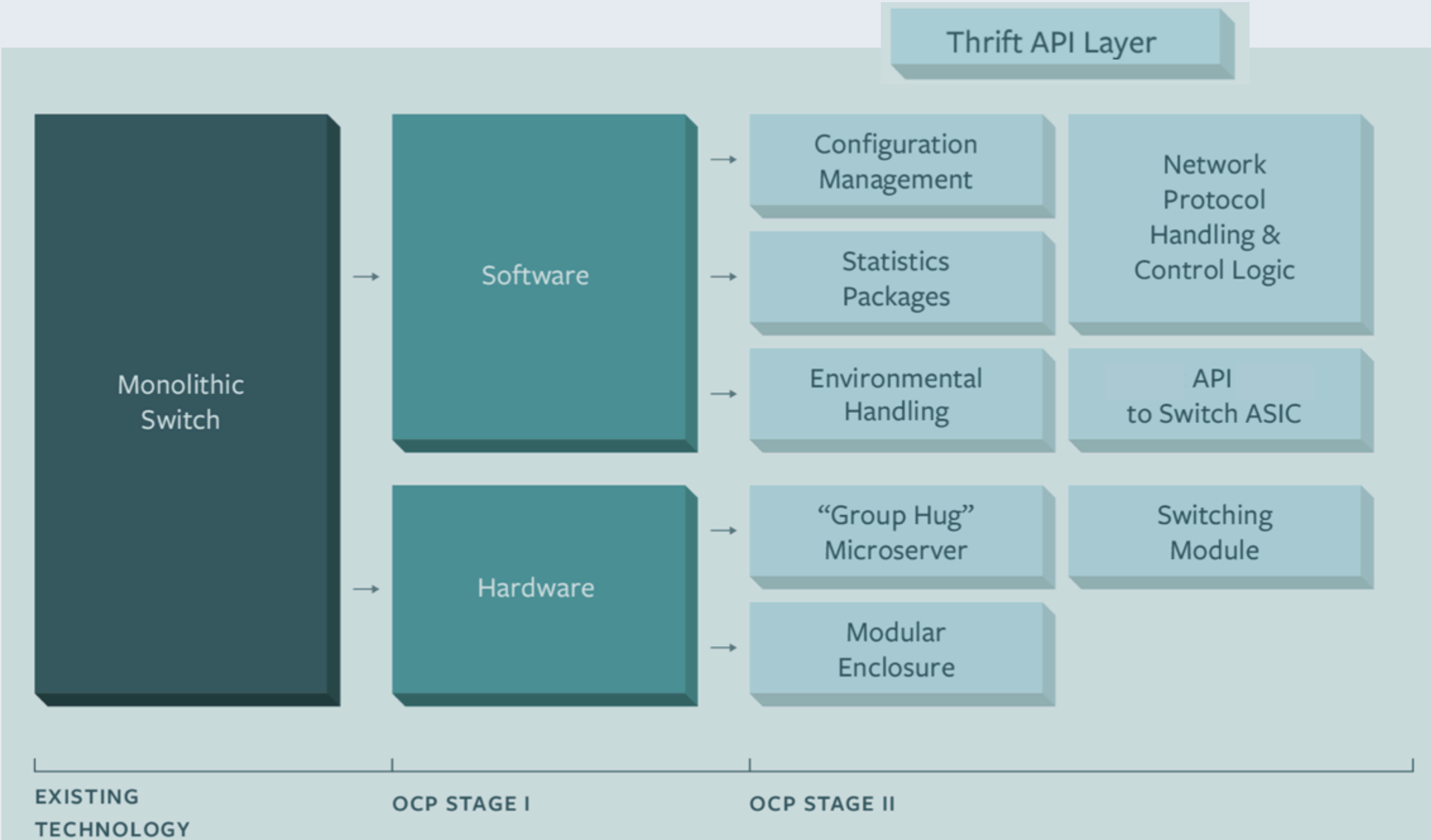


NETWORK

Wedge

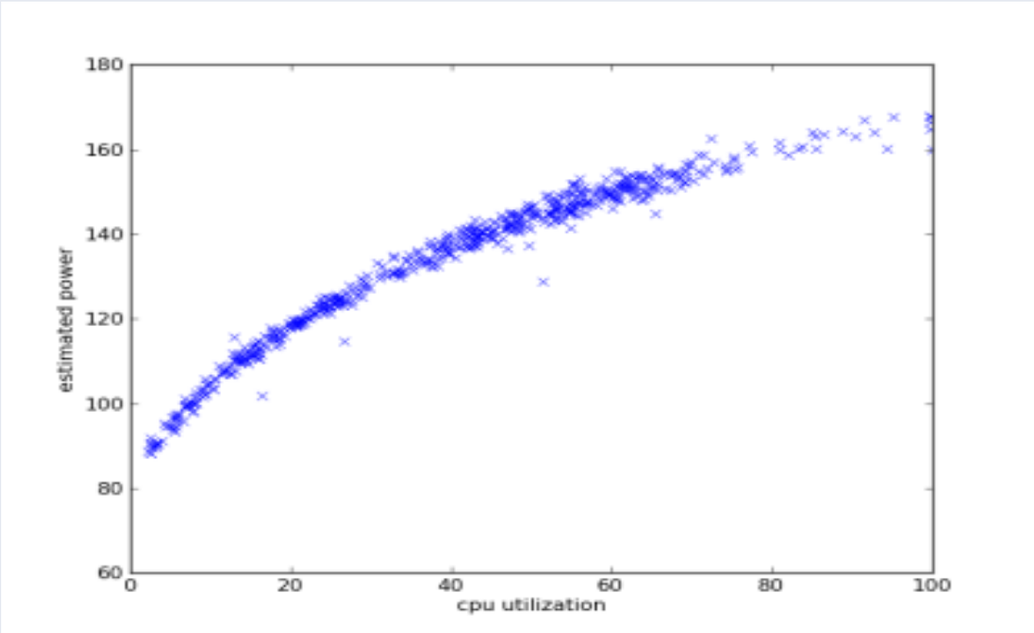

Open Compute "Group Hug"
Micro Server
40Gb switching ASIC
Commercially available
Sixteen 40Gb network ports
spaced for optimal airflow
Dual power supplies
with AC and DC options
Fans
Simple enclosure
optimized for efficient cooling

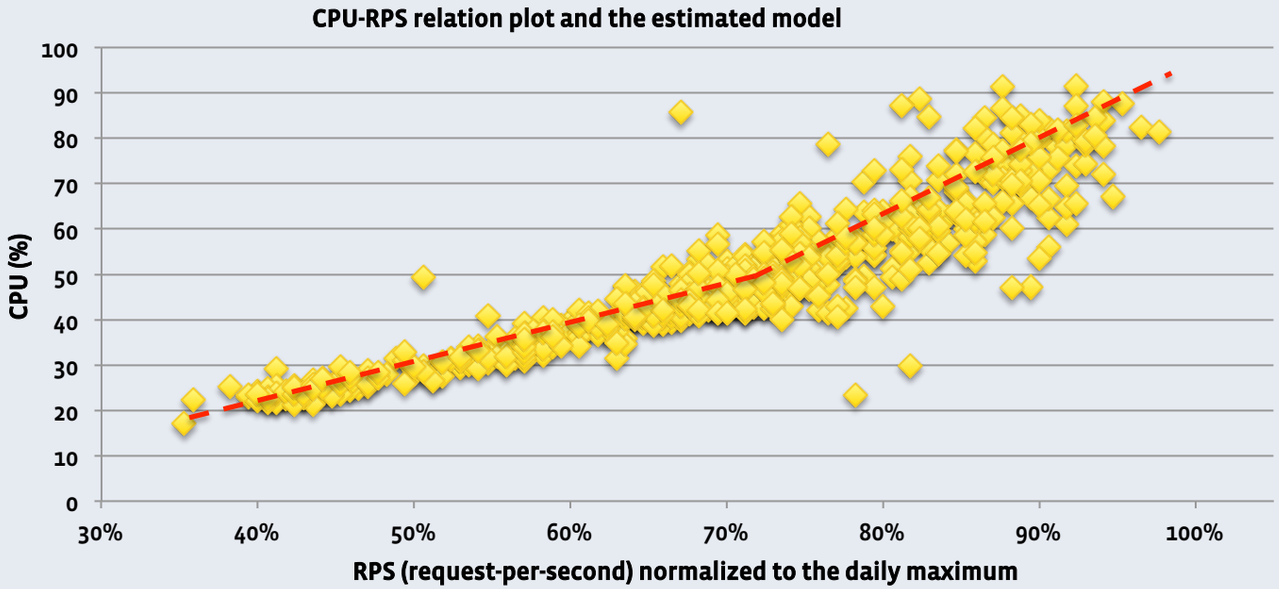
FBOSS


Thrift API Layer
Monolithic Switch
Software
Hardware
Configuration Management
Statistics Packages
Environmental Handling
"Group Hug" Microserver
Modular Enclosure
Network Protocol Handling & Control Logic
API to Switch ASIC
Switching Module
EXISTING TECHNOLOGY
OCP STAGE I
OCP STAGE II

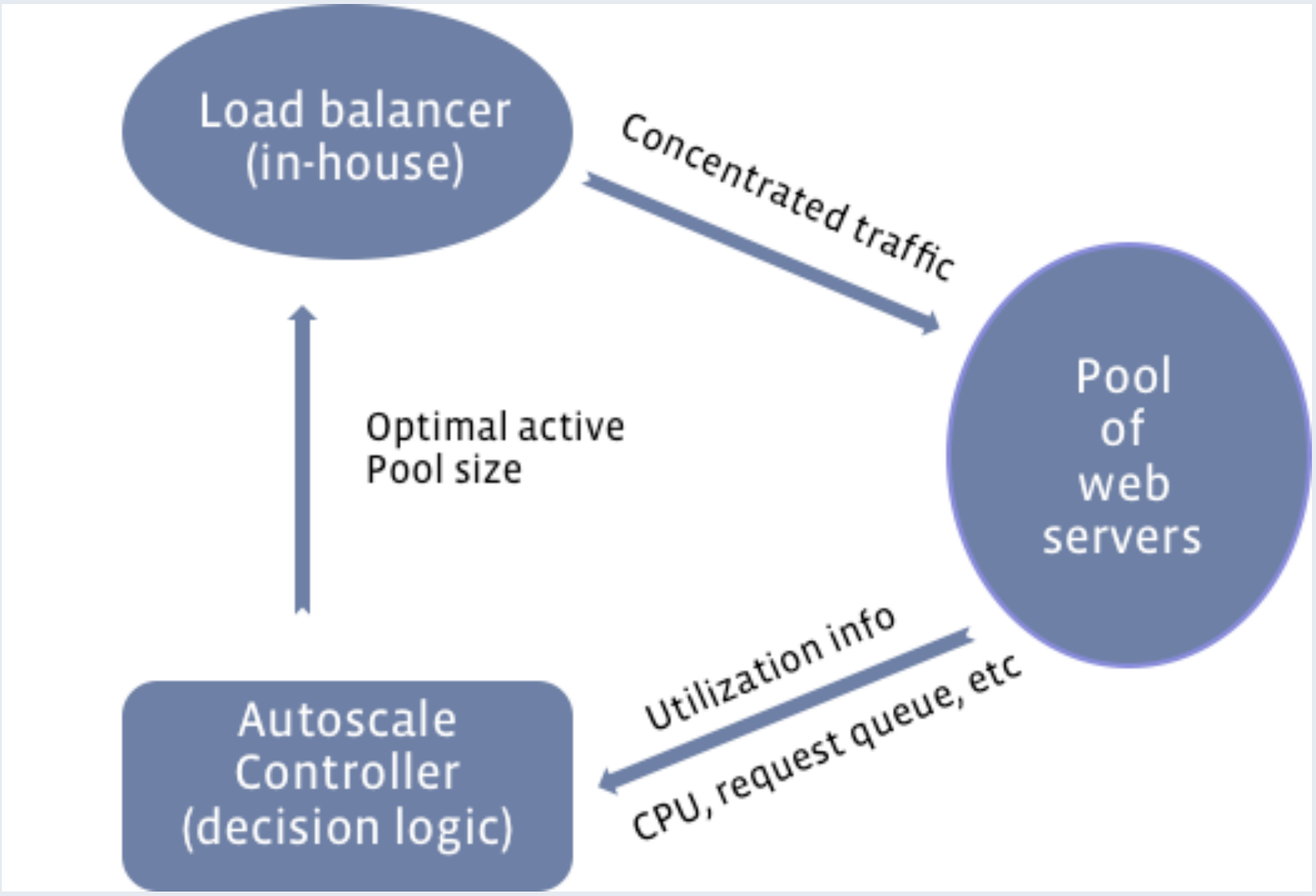
APPLICATION

Power stages

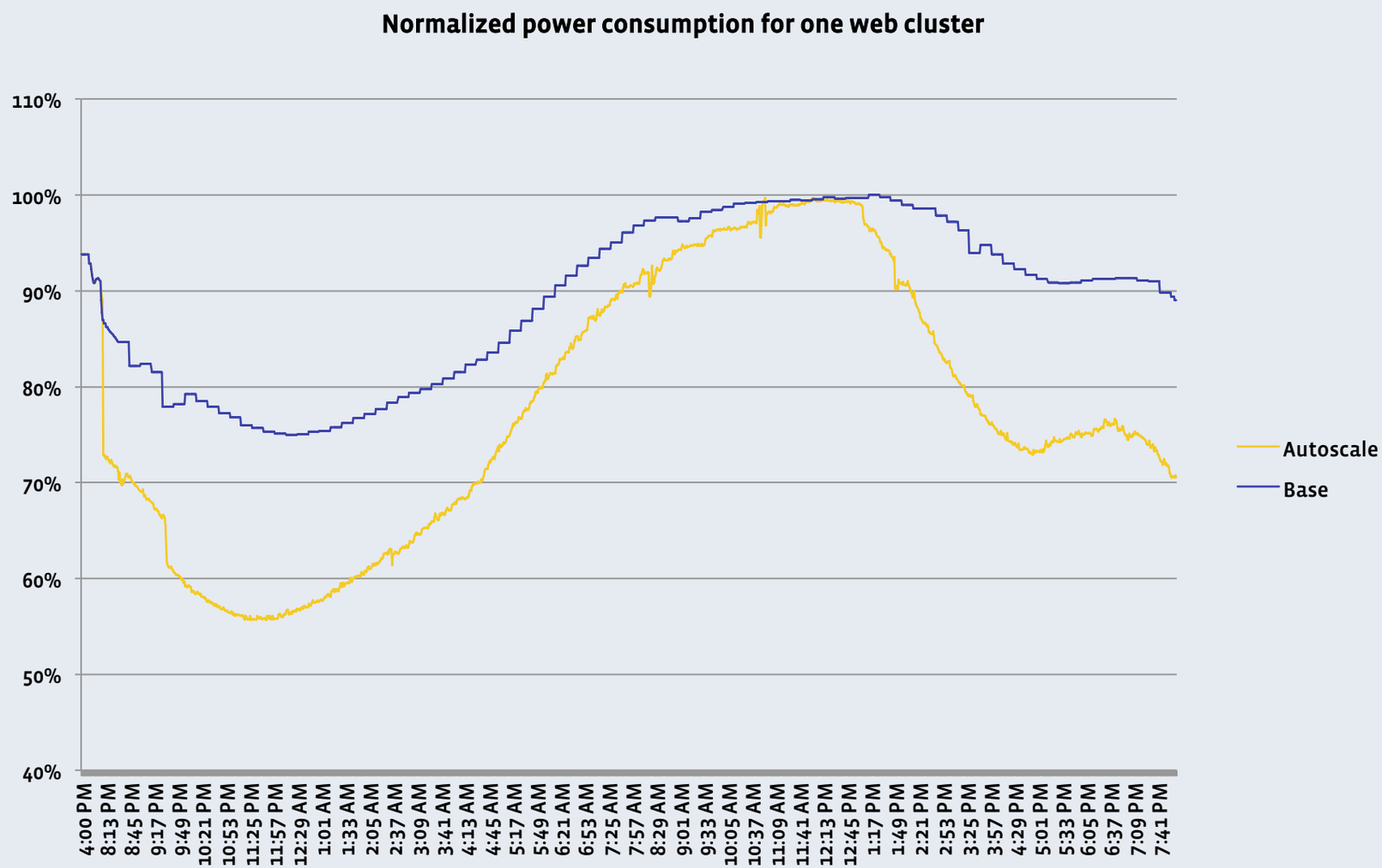




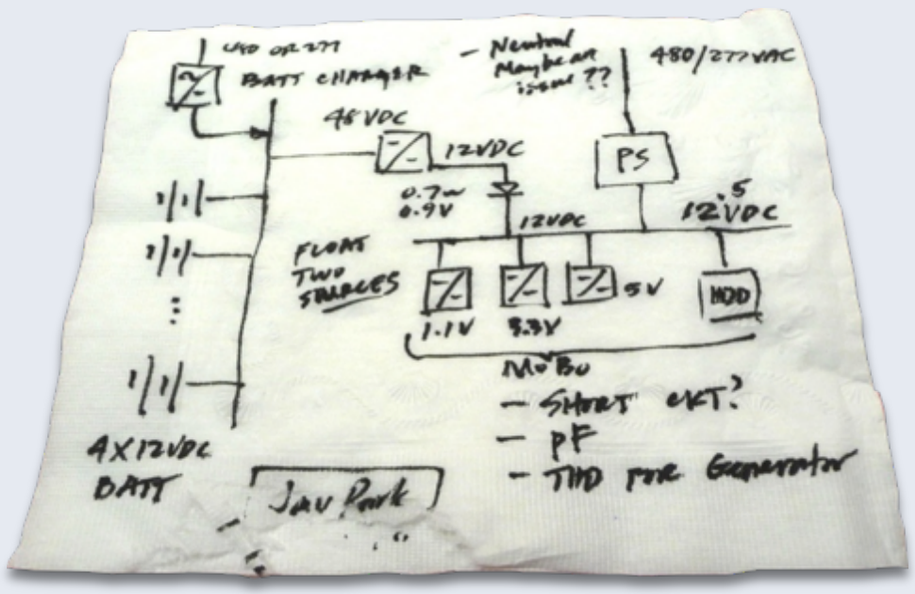
Autoscale: energy efficient load balancing



Power savings



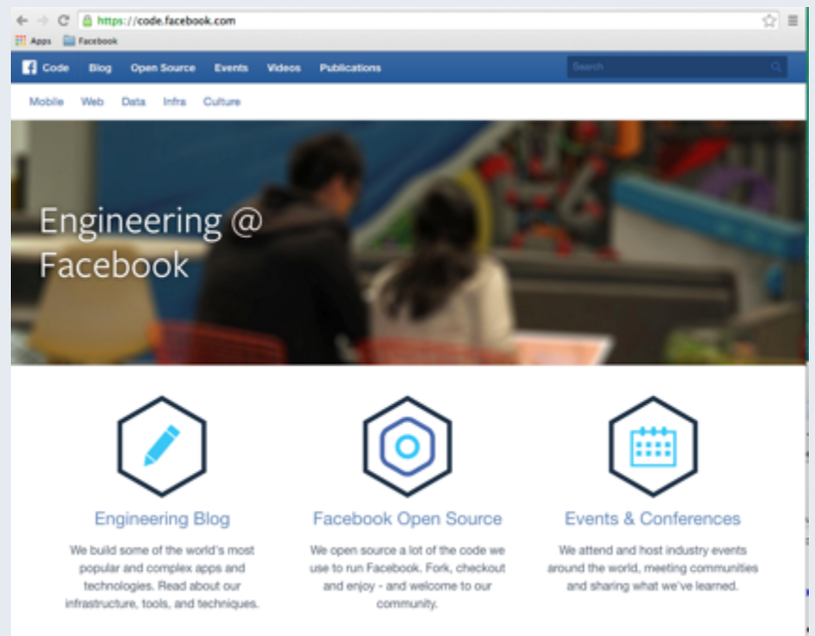
Takeaways



- Take a clean sheet approach
- Efficiency as a design goal
- Look for synergies across the whole cross functional stack
  - data center
  - rack
  - server
  - network
  - application
- Open source

Let’s engage

OPEN
Compute Project


https://code.facebook.com
Code
Blog
Open Source
Events
Videos
Publications
Mobile
Web
Data
Infra
Culture
Engineering @
Facebook
Engineering Blog
We build some of the world's most popular and complex apps and technologies. Read about our infrastructure, tools, and techniques.
Facebook Open Source
We open source a lot of the code we use to run Facebook. Fork, checkout and enjoy - and welcome to our community.
Events & Conferences
We attend and host industry events around the world, meeting communities and sharing what we've learned.

facebook

(c) 2009 Facebook, Inc. or its licensors. "Facebook" is a registered trademark of Facebook, Inc.. All rights reserved. 1.0