Washington Systems Center - Storage

Accelerate with IBM Storage: Looking at the Economic Value of Response Time

Matt Key
FlashSystem SME, Washington Systems Center

© Copyright IBM Corporation 2018.

Performance from Processors

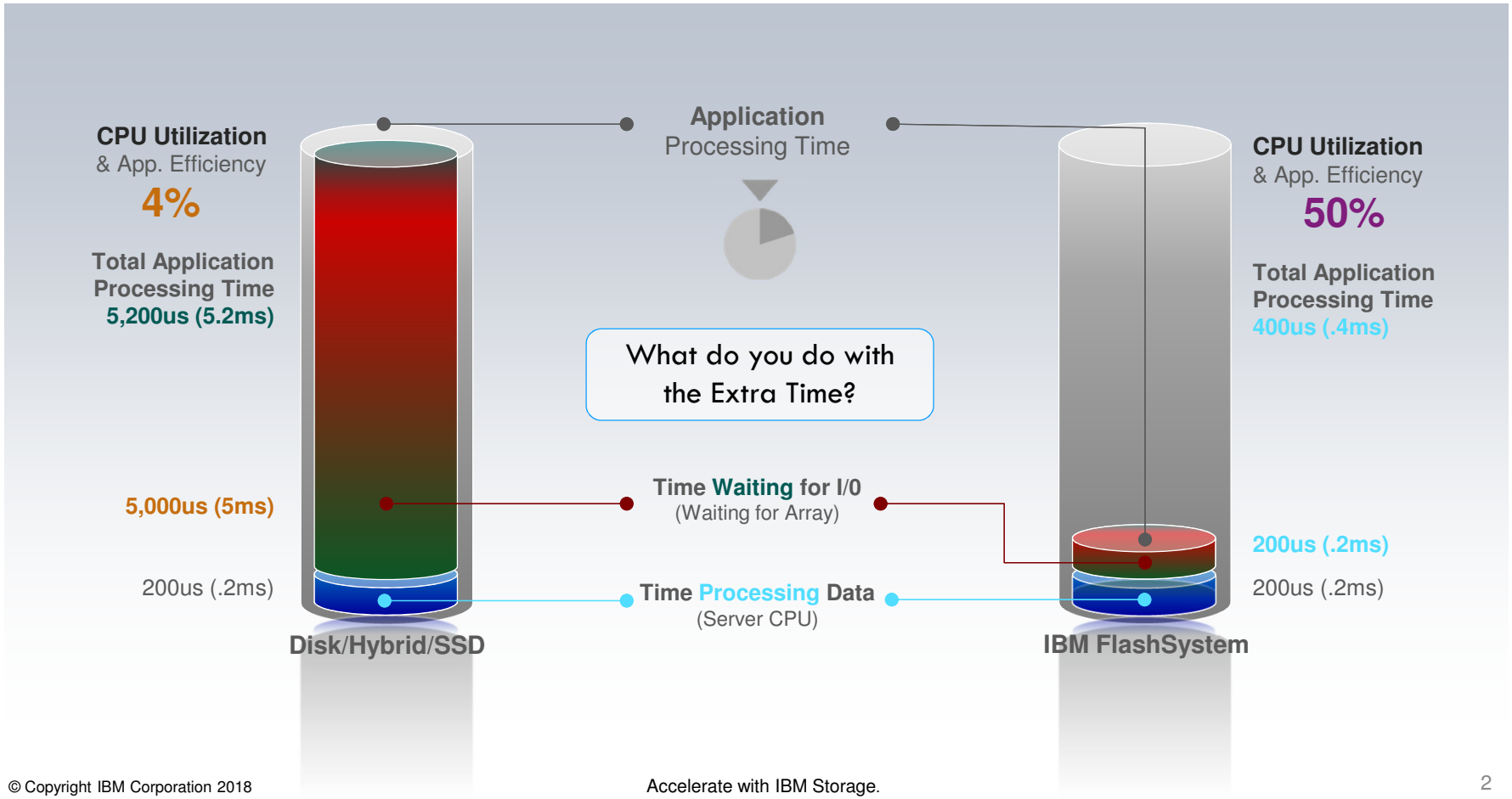
Capacity from Storage

Reality

© Copyright IBM Corporation 2018

The impact of Flash

Flash differentiates namely by serving data ***QUICKER***.



© Copyright IBM Corporation 2018

Benefits of All-Flash

Enable New Workloads

End user experience

Data Mobility

Data Management

Performance

Reduced Infrastructure

Processor Efficiency

Expanded App Services

Storage Longevity

Storage Licensing

Automation

Data Efficiency

Power & Cooling

High-end Virtualization

Lower Admin Stress

Data Reduction

Performance Isolation

Stop Gap

Software Licensing

Time to Results

Recovery

Capacity Density

Productivity

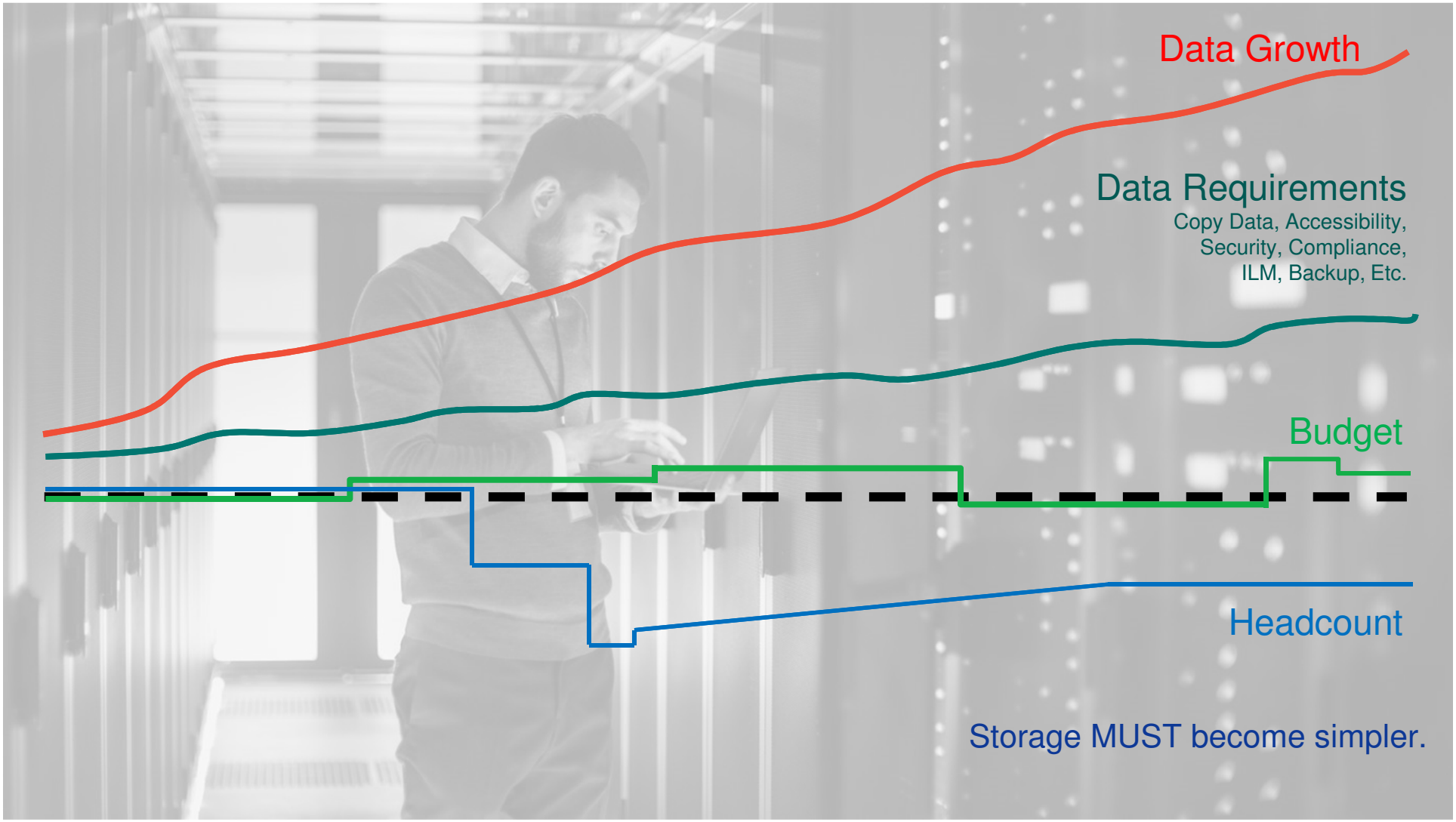
End user efficiency

Server Density

Server Scalability

© Copyright IBM Corporation 2018

Why WE need All-Flash



© Copyright IBM Corporation 2018

What's Holding Flash Back

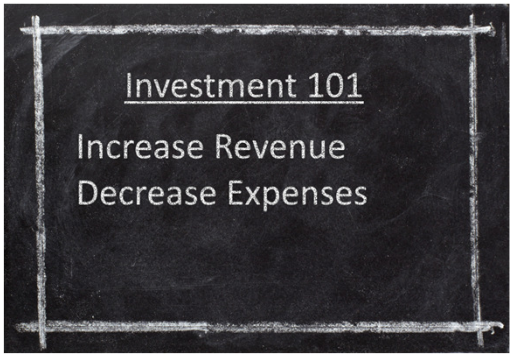
© Copyright IBM Corporation 2018

Champion the All-Flash Momentum

Highlighting how technology provides one of the greatest potential returns in business

Talent | **Data** | **Software Efficiency**

If you want to show value, impact the areas that cost more than storage.



These three areas where Flash storage has tremendous value, has recently expanded. It was once people/talent and licensed cores. Now we acknowledge that data is the new resource and software efficiency can be tied to entire infrastructure.

© Copyright IBM Corporation 2018

IT as Investment for Time

$35.64

Average Employee Cost per Hour in US[1]

x **39.2** hours

Average Time Wasted on Slow Computers per Year in US[2]

x **68.2**%

Average Percentage of Time Waiting for Data[3]

= **$952.81** WASTED

per Employee Annually on Slow Data

[1]https://www.bls.gov/news.release/pdf/ecec.pdf
[2]http://www.telegraph.co.uk/technology/news/10361881/Britons-lose-five-and-a-half-days-a-year-from-slow-computers.html
[3]IBM, Flash Memory Summit 2014

That's about ~1PB of all-flash capacity for every 1,000 employees.

© Copyright IBM Corporation 2018

The Economic Value of Rapid Response Time

"you ask your computer to do something and hit the enter key, if it answers you back in less than 400 milliseconds, just under half a second, then you will stay glued to that machine for hours. Your eyes may glaze over, but your productivity will soar. You'll be transfixed, mesmerized.

Even a slight deviation back to half a second response time will allow your attention to stray. You'll get up and do the dishes, pick up the remote, watch the game.

But under 400 milliseconds, ah, that's the sweet spot."

Halt and Catch Fire "Close to the Metal", 2014

AS SEEN ON TV

© Copyright IBM Corporation 2018

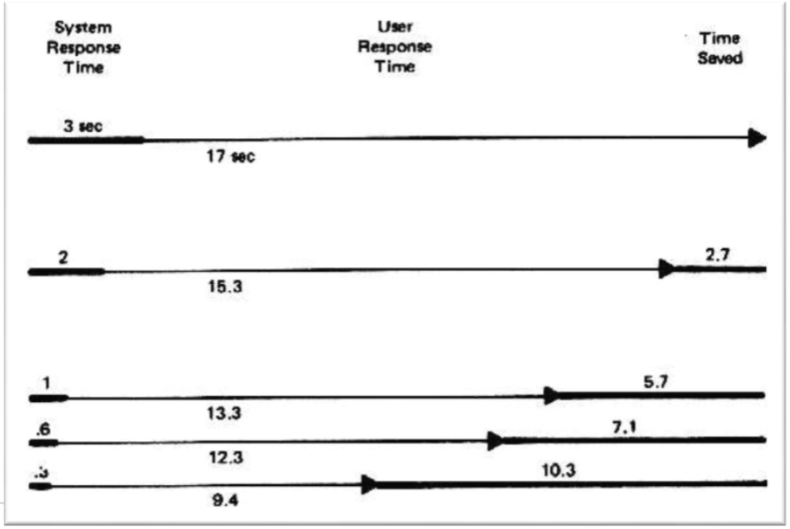
"The Economic Value of Rapid Response Time"

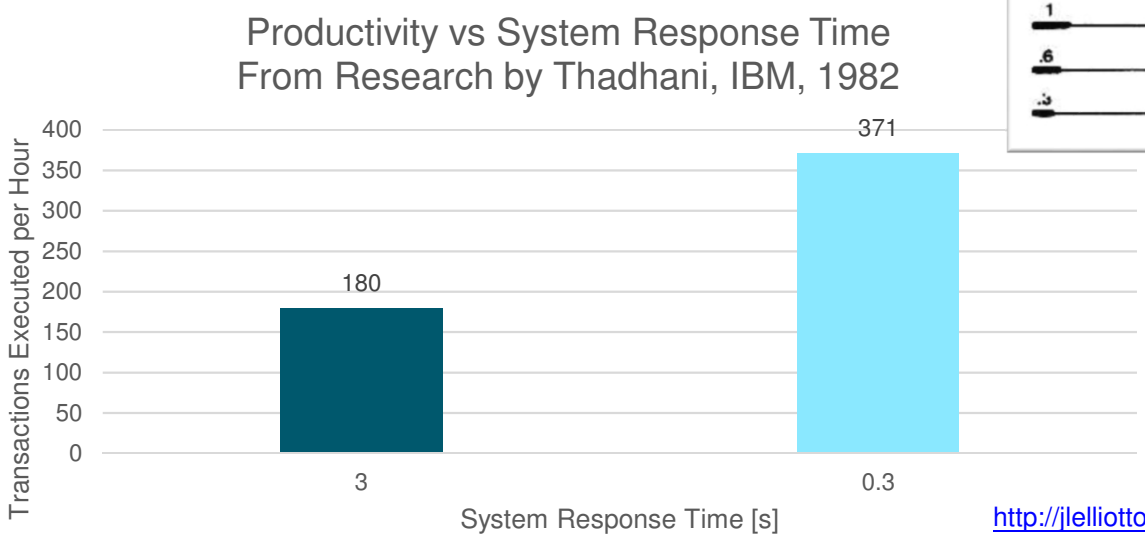
Doherty threshold – System response times should approach 0.4ms, instead of the prior 2 seconds for optimal user response

– IBM Research 1984

"The traditional model of a person thinking after each system response appears to be inaccurate. Instead, people seem to have a sequence of actions in mind, contained in a short-term mental memory buffer. Increases in SRT [system response time] seem to disrupt the thought processes, and this may result in having to rethink the sequence of actions to be continued."

-Walter J.Doherty, IBM, 1979





http://jlelliotton.blogspot.ca/p/the-economic-value-of-rapid-response.html

© Copyright IBM Corporation 2018

Using Little's Law in both Cause and Effect

$$\frac{Queue\ [I/O]}{Time\ [s]} = Rate\ ["IOPS"]$$

Queue = The number of parallel threads running in the application

Time = The Time it takes for an IO request to be serviced

Rate = Result, typically measured in IOPS or Bandwidth

We measure IOPS and Avg Response time…Solve for Queue
(I/O per Sec) x (Avg Sec per I/O) = Avg I/Os Queued

© Copyright IBM Corporation 2018

Leveraging Little's Law for…

Performance

Common workload @ **5ms**

$$\frac{20\ I/O}{0.005s} = 4{,}000\ IOPS$$

Same workload @ **1ms**

$$\frac{20\ I/O}{0.001s} = 20{,}000\ IOPS$$

Same workload @ **0.5ms**

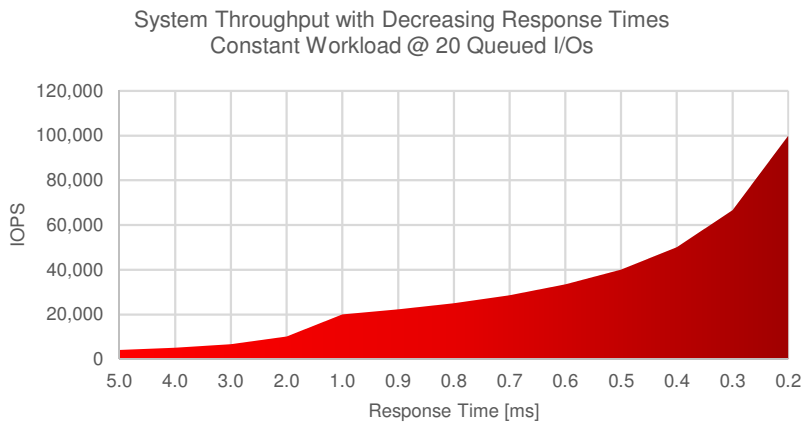
$$\frac{20\ I/O}{0.0005s} = 40{,}000\ IOPS$$

Same workload @ **0.2ms**

$$\frac{20\ I/O}{0.0002s} = 100{,}000\ IOPS$$

By only changing response time, we are able to take existing workloads and finish them in fraction of the time!

Note: As response time approaches 0, impact to performance increases drastically!
(Shave 1/10th of a millisecond off 0.2ms and you'll get 200K IOPS)



AI, machine learning, enriched analytics drive the value of data and are made possible with low-latency I/O.

© Copyright IBM Corporation 2018

Leveraging Little's Law for…

Efficiency

Common workload @ **5ms**

$$\frac{20\ I/O}{0.005s} = 4{,}000\ IOPS$$

By dropping the time wasted in I/O, what use to take 20 threaded I/Os can now be done with one. Query times are now closer to CPU time.

Same work @ **0.250 ms**

$$\frac{1\ I/O}{0.0002s} = 4{,}000\ IOPS$$

Leverage the recovered threads as increased application density OR opportunity to enrich the data for applications.

Application acceleration or point application efficiency was the original driver for flash adoption.

This has been done now for over 10 years. We are now consistently at PB-level flash deployments.

Too many apps…not enough time to value impact…moving on to bigger picture:

© Copyright IBM Corporation 2018

Using Little's Law to Estimate Overall DATA-center Efficiency

- Processors are Processing ☺ or Waiting ☹
  - Data is also on either side
  - While some wait in network, locking, etc most wait is in I/O

$$\frac{CPU\ Queue}{(CPU + Storage)}$$ = % of Time our Data is at Work

- Get on a common metric → Queue
  - Queue of Storage
    - Little's Law → Queue = IOPS * Response Time
  - Queue of Processors
    - Total # of Cores * (% User + % System) Efficiency of Cores
    - 1,000 Cores * 28% = 280 Cores in full use; CPU Queue = 280

© Copyright IBM Corporation 2018

Using Little's Law to Estimate Overall DC Efficiency (Example)

- 1,000 Cores @ 28% Utilization = 280 CPU Queue
- Storage
  - 36,000 IOPS
  - 4.3ms Response Time
  - Queue = $36{,}000\ IOPS\ *\ 0.0043\ s$ = 154 Queued I/Os

$$\frac{CPU\ Queue}{(CPU + Storage)} = \frac{280}{(280 + 154)} = 64\%$$

If we drop response time to 0.5ms…

Queue = $36{,}000\ IOPS\ *\ 0.0005\ s$ = Only 18 Queued I/Os!

$$\frac{280}{(280 + 18)} = 93\%$$

© Copyright IBM Corporation 2018

Recap

1. People, Data, and Infrastructure Efficiency cost more than storage.

2. We only need to save a couple of minutes per day each to cost justify a faster, more efficient, and flexible infrastructure that enables new workloads like AI/ML.

3. AI/ML are such data and I/O beasts that flash is a minimum for most use cases. RAM is the primary focus, but flash is there to supplement and extend the reach of the learning to bigger datasets.

4. Leverage Little's Law to highlight an immediate value of low-latency storage

$$Queue\ [I/O] = Rate\ ["IOPS"] * Avg\ Time\ [s]$$

$$Queue\ [MB] = Rate\ ["MBps"] * Avg\ Time\ [s]$$

© Copyright IBM Corporation 2018

Accelerate with IBM Storage Webinars

The Free IBM Storage Technical Webinar Series Continues in 2018...

Washington Systems Center – Storage experts cover a variety of technical topics.

Audience: Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with IBM Storage webinars, Clients, Business Partners and IBMers are welcome to send an email request to accelerate-join@hursley.ibm.com.

Located in the Accelerate with IBM Storage Blog:
https://www.ibm.com/developerworks/mydeveloperworks/blogs/accelerate/?lang=en

Also, check out the WSC YouTube Channel here:
https://www.youtube.com/playlist?list=PLSdmGMn4Aud-gKUBCR8K0kscCiF6E6ZYD&disable_polymer=true

2018 Webinars:

January 9 – DS8880 Easy Tier
January 17 – Start 2018 Fast! What's New for Spectrum Scale V5 and ESS
February 8 - VersaStack - Solutions For Fast Deployments
February 16 - TS7700 R4.1 Phase 2 GUI with Live Demo
February 22 - DS8880 Transparent Cloud Tiering Live Demo
March 7 - Spectrum Storage Management, Control, Insights, Foundation; what's the difference?
March 15 - IBM FlashSystem A9000/R and SVC Configuration Best Practices
March 27 - IBM FlashSystem A9000/R Technical Update
April 12 - Introducing Spectrum NAS - The Newest Member in the Spectrum Storage Family
April 26 - TS7700 Grid Configuration Changes -- Joins, Merges and Removals
May 8 - DS8880 Technical Update
June 7 – Economic Value of Response Time
June 21 – Deep Dive into Spectrum Scale AFM
Register Here: https://ibm2.webex.com/ibm2/onstage/g.php?MTID=eb2ec98d28a8e2e1d239dca974293b96f
June 28 - Open System Tape, What's New
Register Here: https://ibm2.webex.com/ibm2/onstage/g.php?MTID=e74c4d48c903c2dbff250554ee538c0e4

© Copyright IBM Corporation 2018

Thank YOU!

© Copyright IBM Corporation 2018