Introduction to zProcessor Measurements

Scott Chapman
Enterprise Performance Strategies, Inc.
Scott.chapman@EPStrategies.com

Contact, Copyright, and Trademarks

Questions?

Send email to performance.questions@EPStrategies.com, or visit our website at https://www.epstrategies.com or http://www.pivotor.com.

Copyright Notice:

Trademarks:

Enterprise Performance Strategies, Inc. presentation materials contain trademarks and registered trademarks of several companies.

The following are trademarks of Enterprise Performance Strategies, Inc.: **Health Check®, Reductions®, Pivotor®**

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries: IBM®, z/OS®, zSeries®, WebSphere®, CICS®, DB2®, S390®, WebSphere Application Server®, and many others.

Other trademarks and registered trademarks may exist in this presentation

Abstract

- On the z platform there are more processor measurements than any other computing platform. What are all these processor measurements actually measuring, where do they come from, and how can they be used? During this presentation Scott Chapman will introduce to you the many processor measurements available, and show you how many of them can be used during any processor performance analysis or capacity planning exercise.

Agenda

- CPU Terms
- CPU Capacity and Speeds
- CPU Measurements
- CPU Analysis

EPS: We do z/OS performance…

- Pivotor - Reporting and analysis software and services
  - Not just reporting, but analysis-based reporting based on our expertise
- Education and instruction
  - We have taught our z/OS performance workshops all over the world
- Consulting
  - Performance war rooms: concentrated, highly productive group discussions and analysis
- Information
  - We present around the world and participate in online forums

z/OS Performance workshops available

During these workshops you will be analyzing your own data!

- Essential z/OS Performance Tuning
  - October 3-7, 2022
- WLM Performance and Re-evaluating Goals
  - September 12-16, 2022
- Parallel Sysplex and z/OS Performance Tuning
  - August 8-12, 2022
- Also… please make sure you are signed up for our free monthly z/OS educational webinars! (email contact@epstrategies.com)

Like what you see?

- The z/OS Performance Graphs you see here come from Pivotor™

- If you don't see them in your performance reporting tool, or you just want a free cursory performance review of your environment, let us know!
  - We're always happy to process a day's worth of data and show you the results
  - See also: http://pivotor.com/cursoryReview.html

- We also have a free Pivotor offering available as well
  - 1 System, SMF 70-72 only, 7 Day retention
  - That still encompasses over 100 reports!

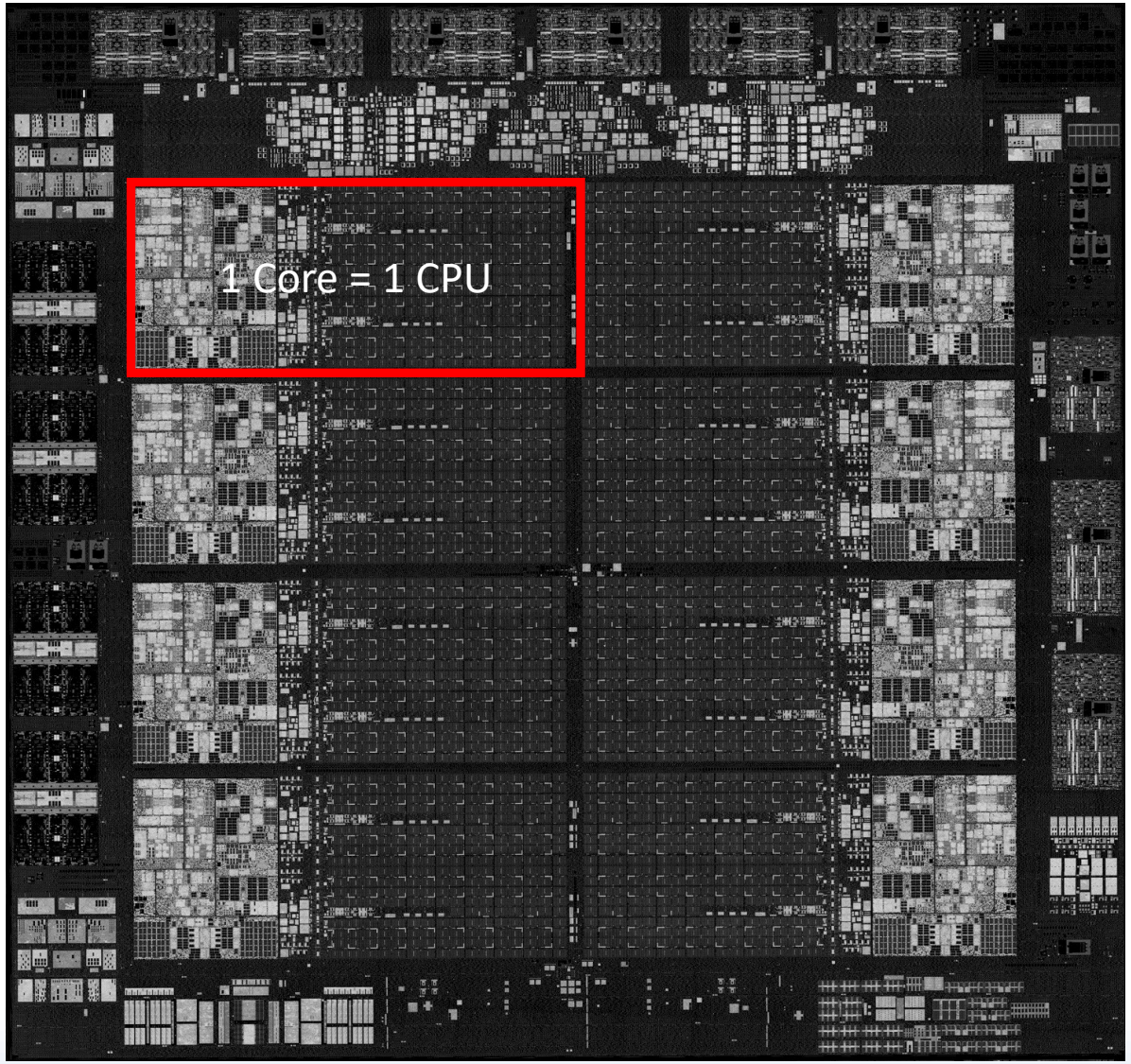
EPS presentations this week

What	Who	When	Where
Introduction to z Processor Measurements	Scott Chapman	Mon 10:30	Cumberland L
Introduction to WLM Management of CICS and IMS Workloads	Peter Enrico	Mon 14:15	Cumberland K
Introduction to the WLM	Scott Chapman	Tue 13:00	Cumberland AB
z/OS WLM – Revisiting Goals over Time	Peter Enrico	Wed 8:00	Cumberland L
Top WLM Mistakes and Questions	Peter Enrico Scott Chapman	Wed 13:00	Cumberland AB

CPU Terms

CPU Characterizations

- Generically, a CPU is a core on a chip
- CPUs can be “characterized” for use with specific work:
  - GP (General Purpose, aka CP)
  - zIIP
  - IFL
  - ICF
  - SAP
  - IFP
  - Spare (not characterized)
- All are physically the same!



Measuring different characterizations

- Because they're all the same, you can generally assume that which holds true for one characterization holds for another
- But what work is allowed to run on each characterization varies
- While the measurements all derive from the same place, sometimes:
  - The measurements are expressed differently
  - The measurements that we care about might be different

SMT

- Simultaneous Multi-Threading (SMT) allows 2 threads (processes) to use the same core at the same time
- May be enabled on:
  - zIIP
  - IFL
  - SAP
- Note that SMT is not allowed on the GPs—primarily because of software licensing concerns
  - Software costs often based on CPU measurements (one way or another)
  - SMT makes CPU measurements more variable

CPU terminology can be confusing

MIPS

Percent Utilization

Appl %

MSUs

Percent Busy

CPU Using

SUs

Workload %

MVS Busy

CPU Seconds

Most of the raw SMF measurements are going to be CPU seconds or SUs

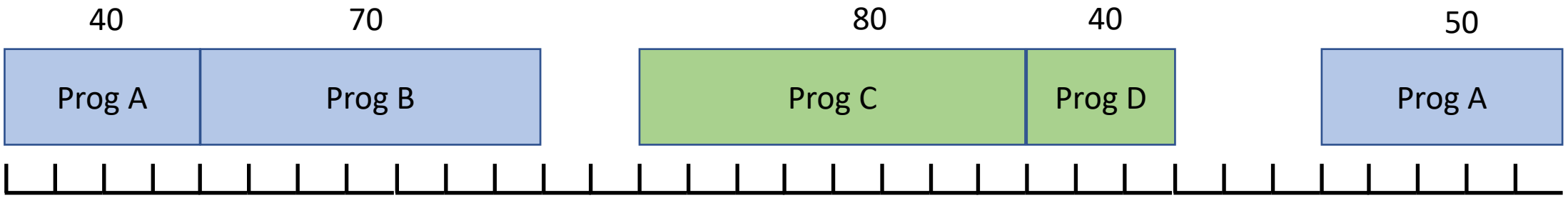
MT1ET

CPU Time

- CPU time = total time that a CPU has spent performing work for task
  - Time that a workload is dispatched to a CPU
- z/Architecture provides CPU timers with nominal resolution of 1 µs
  - Yes, that's one millionth of a second, although the times aren't usually externalized to that precision
- Instruction EXTRACT CPU TIME (ECTG) can be used by problem-state programs to determine the amount of CPU time consumed by the current task
- When a CPU is interrupted to process something else, the CPU timer is readjusted once the interrupted task is dispatched again
- CPU timers are not dependent on the time-of-day clock because (e.g.) the time of day clock may be steered to remain in sync with a time source
  - One of the reasons why use of system time for performance analysis is problematic

What is using the CPU?

No SMT!



Time (on order of microseconds)

So in this example:
ProgA consumed 40 + 50 = 90µs of CPU time
ProgB consumed 70µs
Blue LPAR consumed 40+70+50 = 160µs

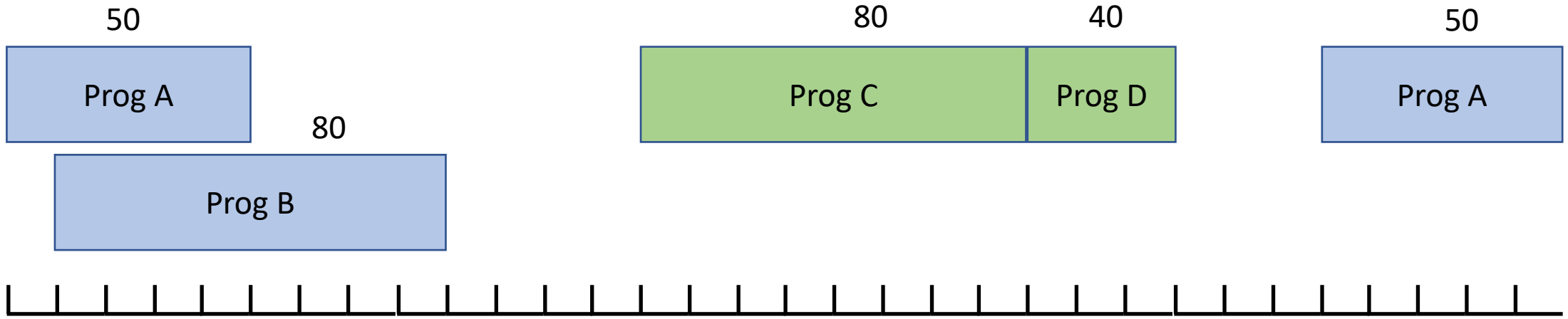
ProgC consumed 80µs
ProgD consumed 40µs
Green LPAR consumed 120µs

The key point here is that (absent SMT) only one program/task from one LPAR can be using the CPU at a time!

Even though at the macro scale it feels like things are sharing the CPU, at a microsecond by microsecond basis, they are not!

What about SMT?

With SMT on Blue



Time (on order of microseconds)

- Note that A & B both spent longer on the CPU because they were contending for the same on-core resources.
- But the Blue LPAR's total usage of the processor has dropped from 160µs to 140µs
- But Blue's programs A + B = 180µs!
- To deal with this, the CPU times reported for programs A an B will be in MT1ET – Mult-Threading 1 Equivalent Time (but don't expect it to be exactly equivalent!)
  - So in the actual records we might see 92µs for Prog A and 78µs for Prog B (maybe)
- Remember: SMT only impacts zIIPs, IFLs, SAPs

SMT makes things more complicated, so let's (mostly) ignore it!

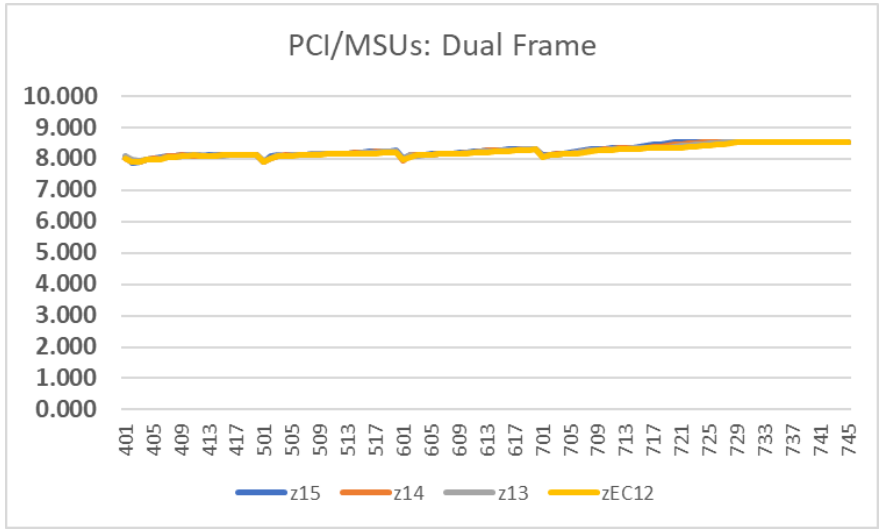
CPU Capacity and Speed

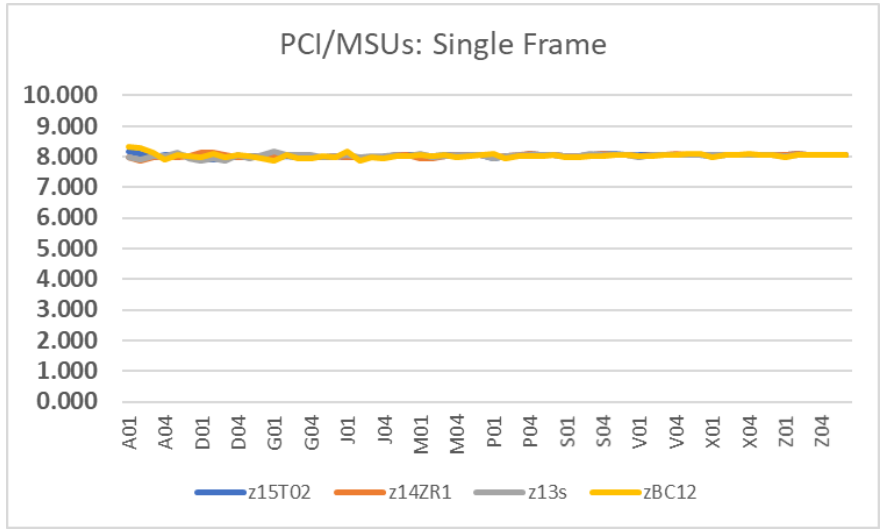
CPU Capacity

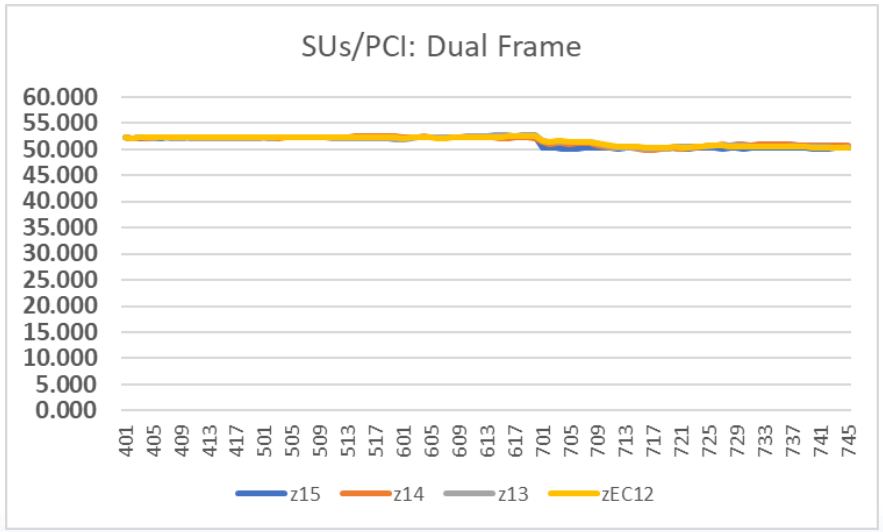
- In this presentation, capacity = CPU capacity
- Capacity = How much of a certain type of work can be done per unit of time
- There are three terms we use to express the capacity:
  - MIPS (or PCI in IBM-speak)
  - MSUs
  - SU/Sec
- IBM publishes PCI, MSU, and SU/sec ratings for each machine
  - All are derived from the same IBM test workload results (LSPR)
  - All are pretty much the same, just different scales
    - I.E. there's a more or less constant relationship between PCI, MSU and SU/sec
- Other vendors publish MIPS ratings for machines
  - May provide more nuance than IBM's ratings

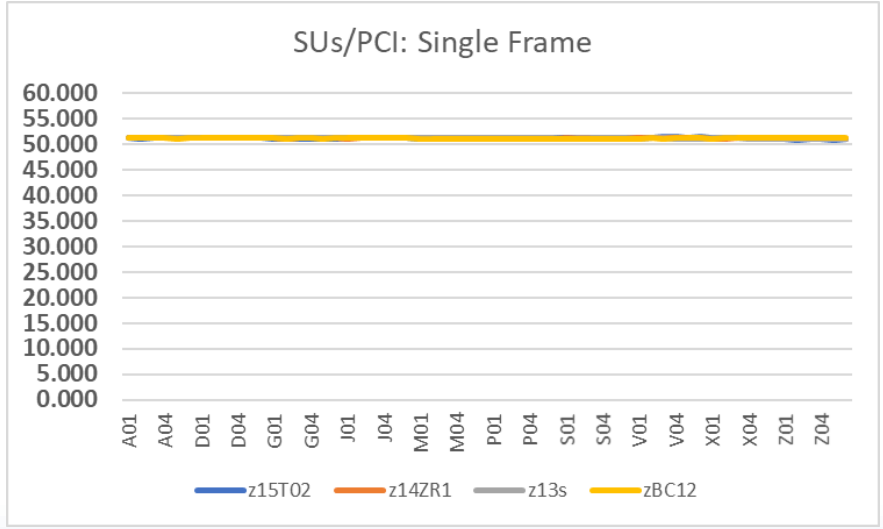
Relating SU/sec, MSUs, PCI

- There are some minor variations in the ratios due to how they do the rounding and likely due to marketing goals.
- MIPS and MSUs are primarily used in pricing software
- SUs/sec used internally by z/OS for certain work management functions









https://www-01.ibm.com/servers/resourcelink/lib03060.nsf/pages/lsprITRzOSv2r3#z15

z15
(System z9 2094-701 = 1.00)

Processor	#CP	PCI**	MSU***	Low*	Average*	High*
8561-401	1	267	33	0.48	0.48	0.45
8561-402	2	512	65	0.95	0.91	0.84
8561-403	3	750	95	1.40	1.34	1.22

https://www-01.ibm.com/servers/resourcelink/lib03060.nsf/pages/srmindex#z15

z15

Processor	STIDP Type	STSI Model	#CP	SU/SEC	SRMsec/RealSec
8561-401	8561	401	1	13937.2822	324.1228
8561-402	8561	402	2	13344.4537	324.1228
8561-403	8561	403	3	13039.9348	324.1228

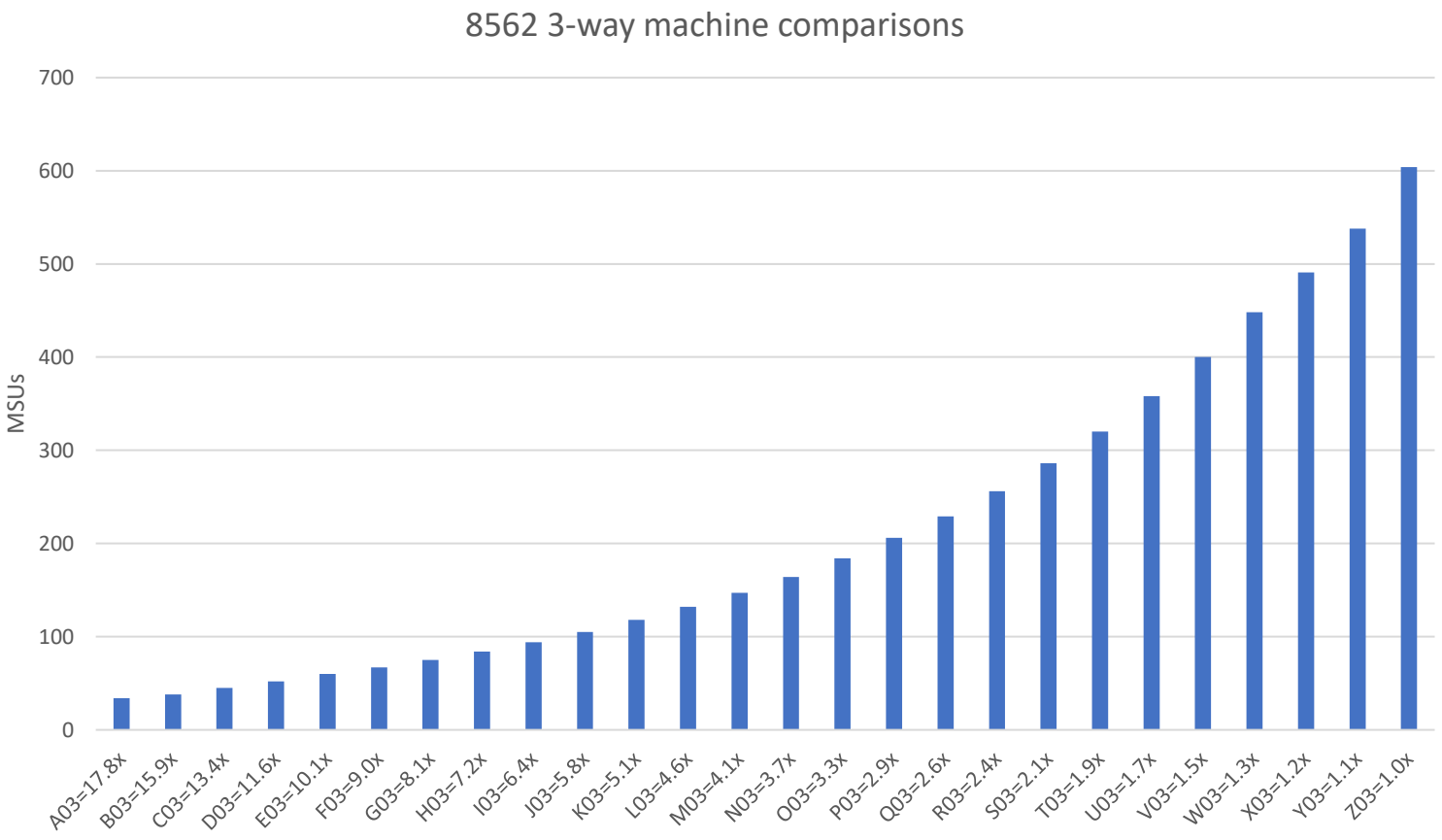
- Note MSU/PCI/MIPS ratings are for the overall machine capacity
- SU/sec is a rating for a single CP on a particular machine (contention results in less total capacity per CP at larger n-ways)
- Do not be deceived into thinking that because SU/sec has many more digits that it's somehow more accurate: it's not!

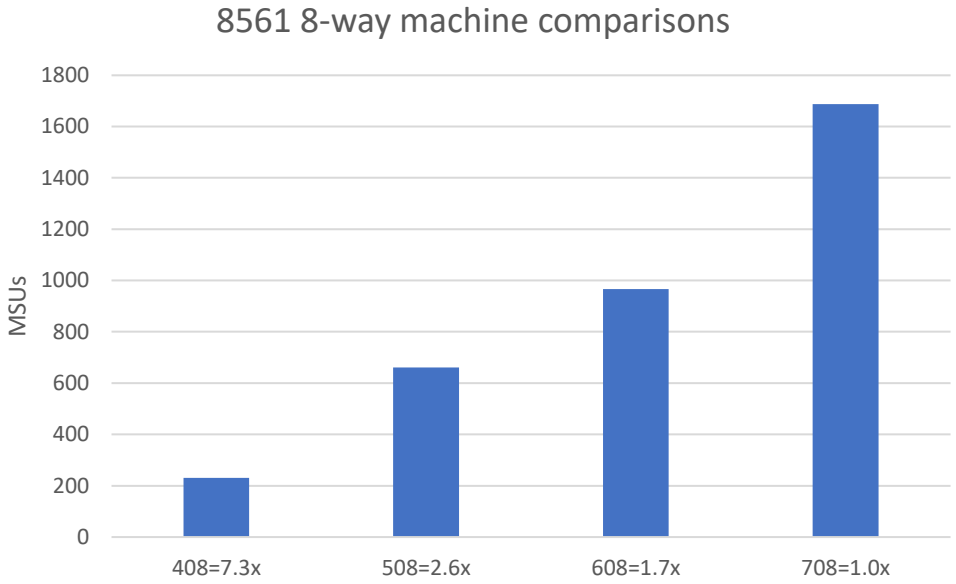
Fast vs. Slow CPs

- You may hear people talk about sub-capacity (or “knee-capped”) processors
  - Or the converse: “full-speed” engines
- Because software costs generally are related to capacity, it may be necessary to buy less capacity
  - But even a single full-speed engine is more capacity than small sites may want!
- Hence, you can buy your machine with “slower” GP CPUs
  - All other CPUs run full speed
  - Clock speed is physical same across all CPUs, but sub-cap GPs will waste a certain amount of time, effectively appearing as if they are running at a slower clock speed
- Note that all characterizations except GPs always run “full-speed”

Comparing engine “speeds”

- How much faster are full speed engines: potentially a lot!





CPU Measurements

Why are you interested in CPU measures?

- Capacity
  - How much of our machine capacity are we using?
  - Are we reaching some capacity limit?
  - How much of the machine is a particular workload using?
- Performance
  - Is the CPU usage by a particular workload changing?
  - Is a particular workload CPU-limited?
- Specific Problem Determination
  - What caused the spike in utilization from 10:00-10:02?

SMF has lots of CPU measurements!

SMF 30	SMF 79	SMF 98	SMF 120
SMF 42	SMF 84	SMF 99	SMF 121
SMF 70	SMF 89	SMF 101	SMF 1153
SMF 72	SMF 96	SMF 110	
SMF 74	SMF 97	SMF 113	

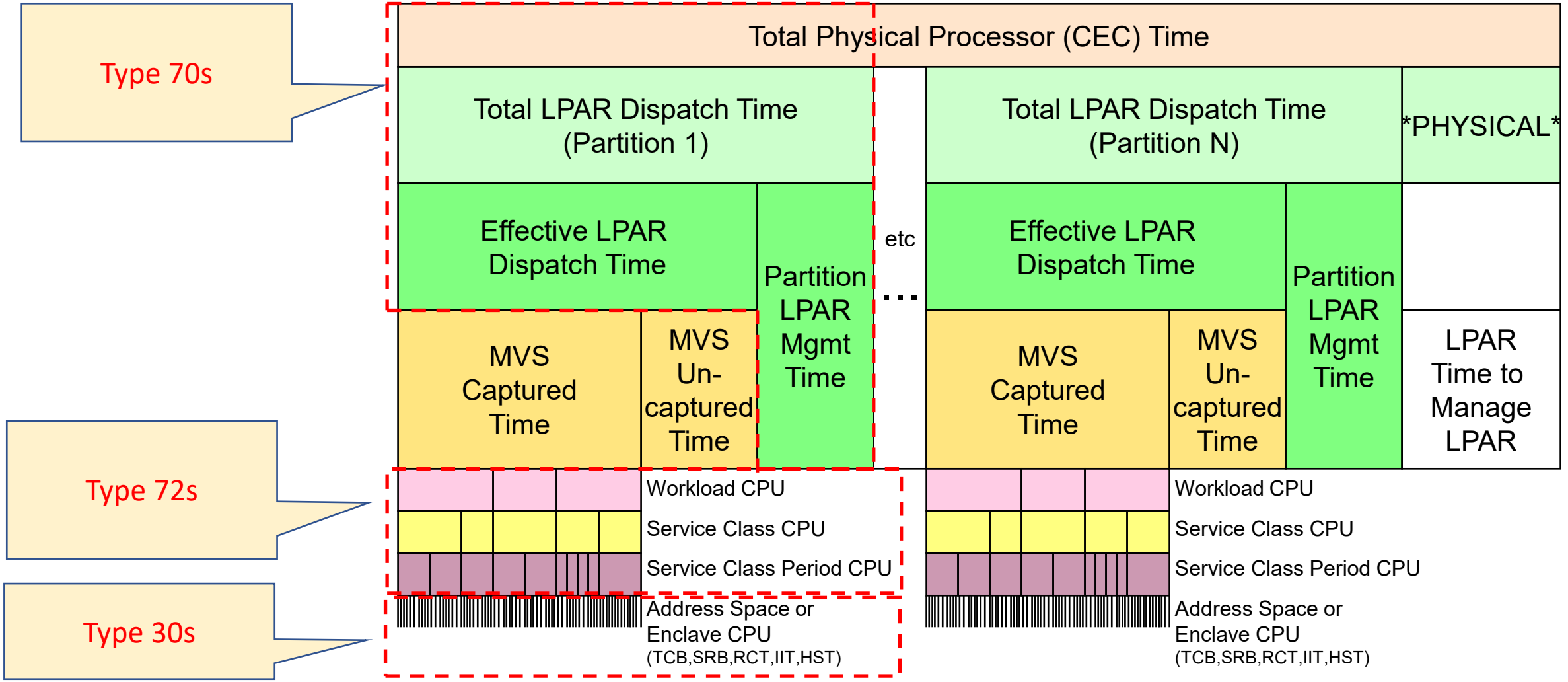
This list is incomplete!!

And we're not going to talk about all of these!

Primary Sources of CPU Measurements

- SMF 70 – LPAR
- SMF 72 – Workload (service class / report class)
- SMF 30 – Address Space
  - End of step (subtype 4) and end of job (subtype 5) records contain totals
    - Often used for chargeback and application attribution
  - Interval records (subtypes 2 and 3) record activity for just 1 interval
    - Can be used to find top consumers within a Service Class
- Recommendations:
  - Set RMF to sync with SMF interval
  - Set SMF interval to no more than 15 minutes
- SMF 99 – Has CPU measurements at sub-minute intervals

Breaking Down CPU Consumption

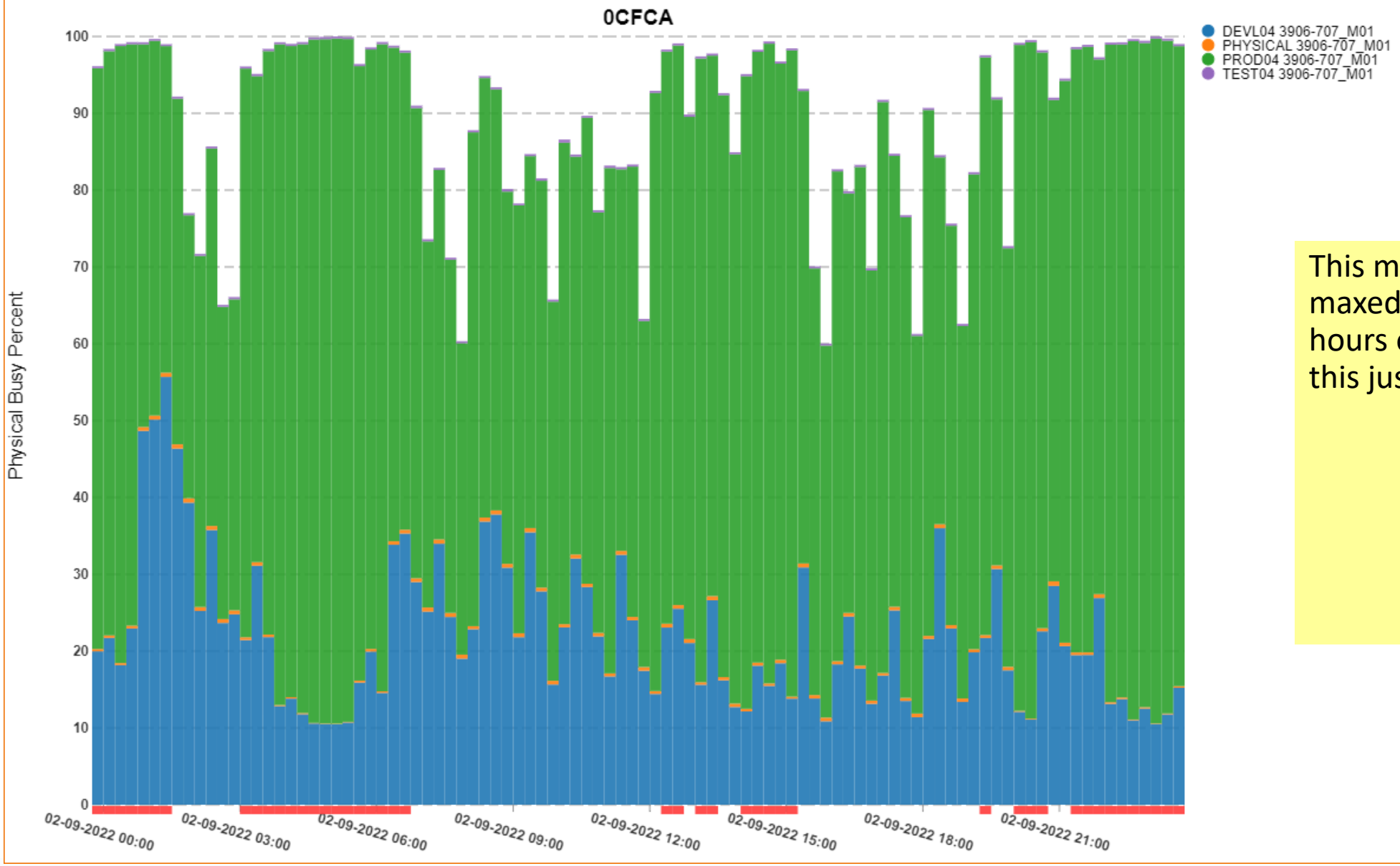


CPU Analysis

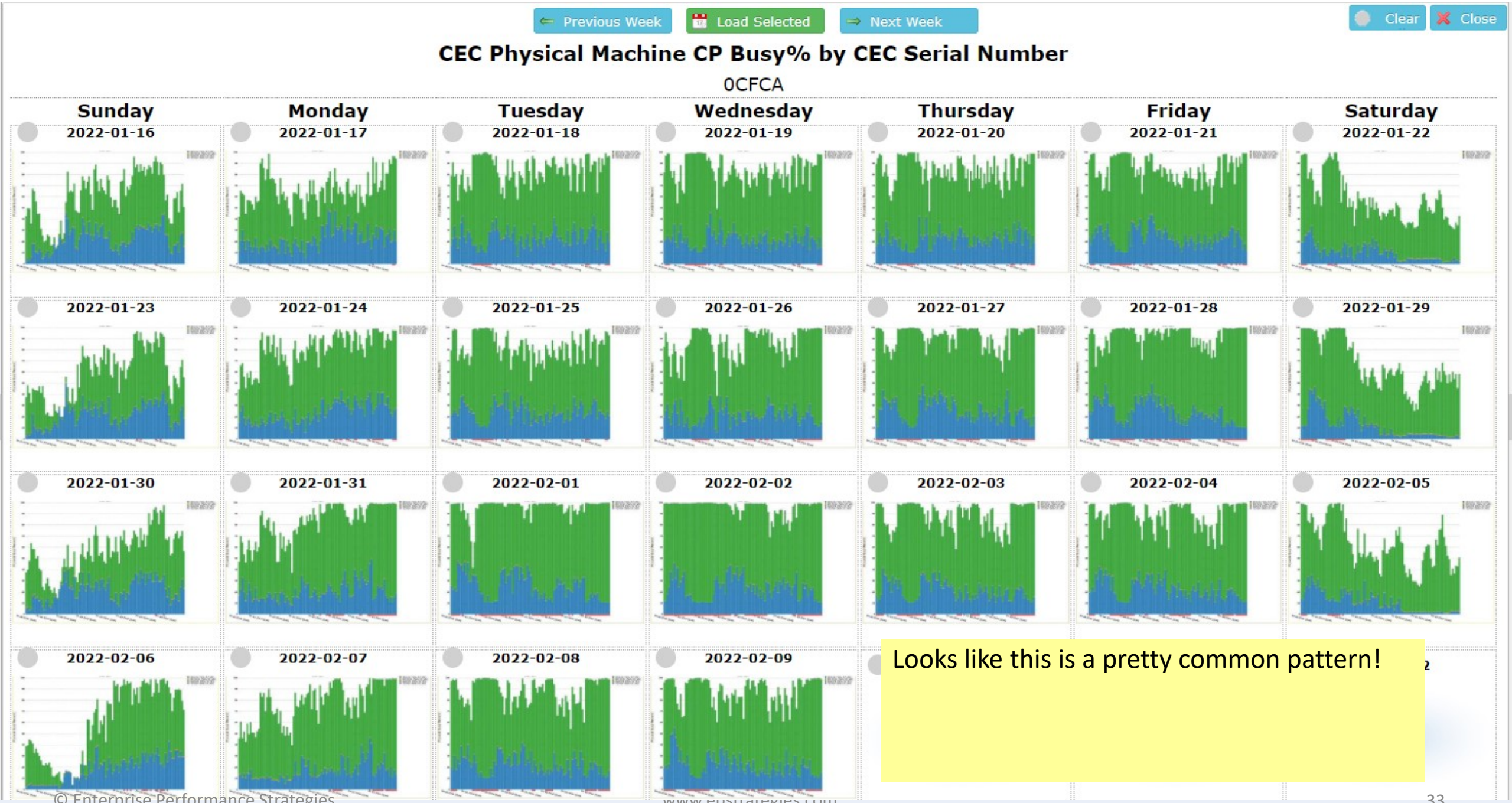
Capacity Management

- How busy is the machine overall is the number one question people want answered
- Generally, for capacity planning want to look at this over time to understand where your problem spots commonly are
- You may run your machine at 100% busy: is that a problem?
  - If not, how do you tell when you're going to run into a problem?

CEC Physical Machine CP Busy% by CEC Serial Number



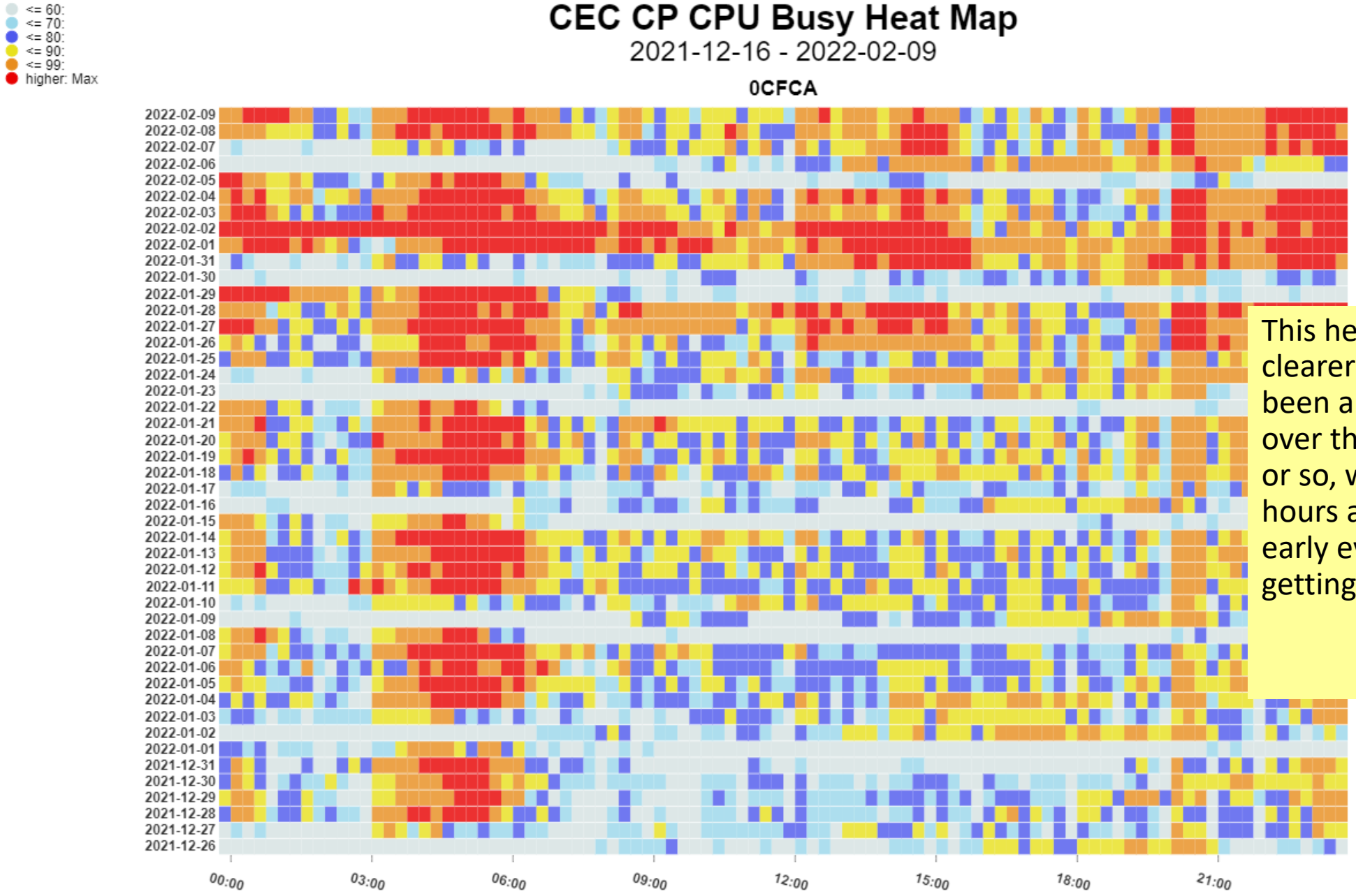
This machine is basically maxed out for multiple hours of the day. But is this just an unusual day?



Looks like this is a pretty common pattern!

CEC CP CPU Busy Heat Map

2021-12-16 - 2022-02-09

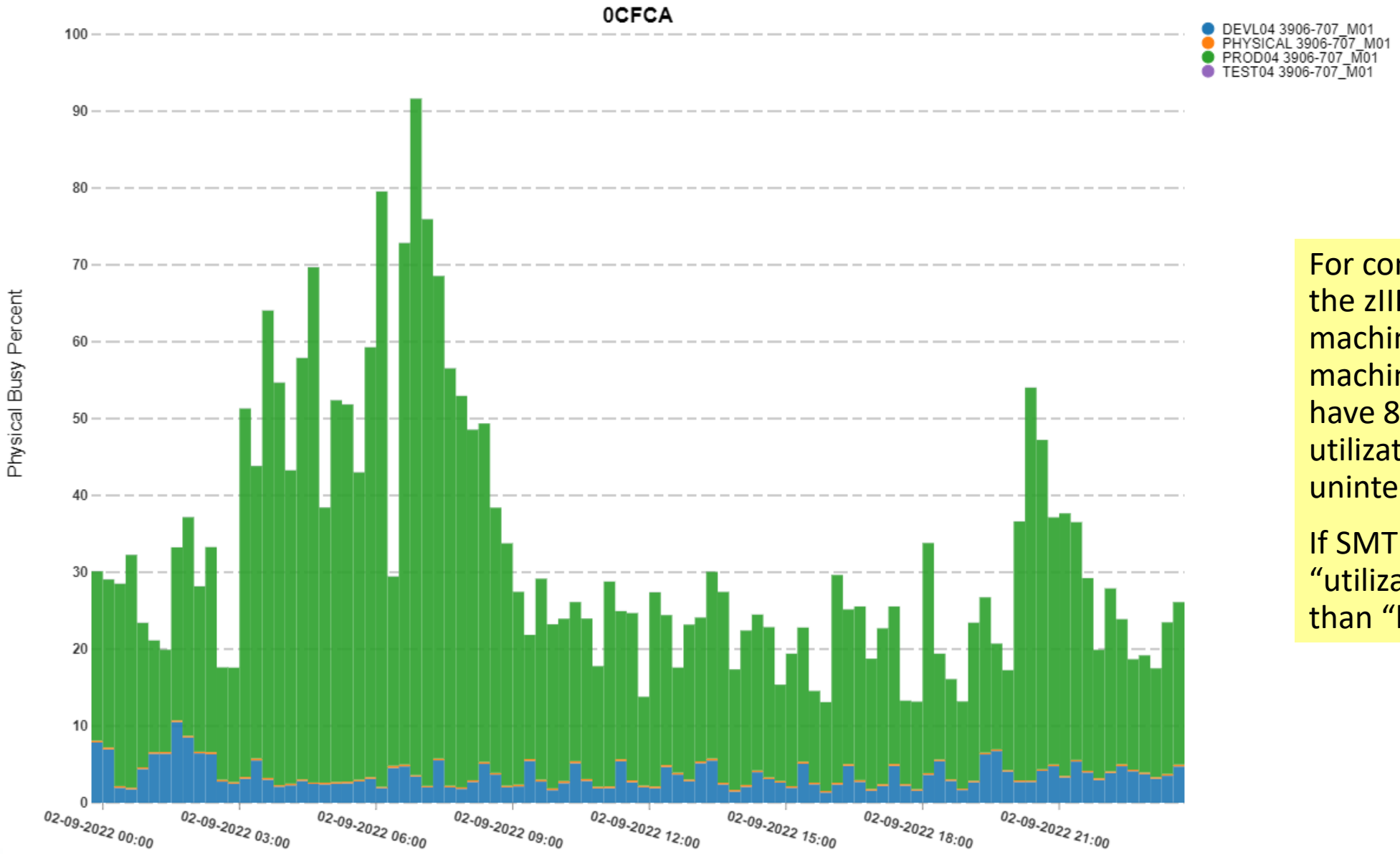


This heat chart makes it clearer that there has been a bit of a change over the past 2.5 weeks or so, with the daytime hours and some of the early evening both getting busier.

CEC Percent Busy

- "How busy is the machine?" comes from the SMF 70 data and sums up the time that the CPUs were dispatched to an LPAR.
- "Physical" series is related to PR/SM management time for managing the LPARs
- Typically one does this for each processor pool
- Note that for GPs, busy = utilization, but for zIIPs with SMT, what constitutes the utilization % is a bit more difficult
  - But for capacity planning purposes, safest to just use "busy"!

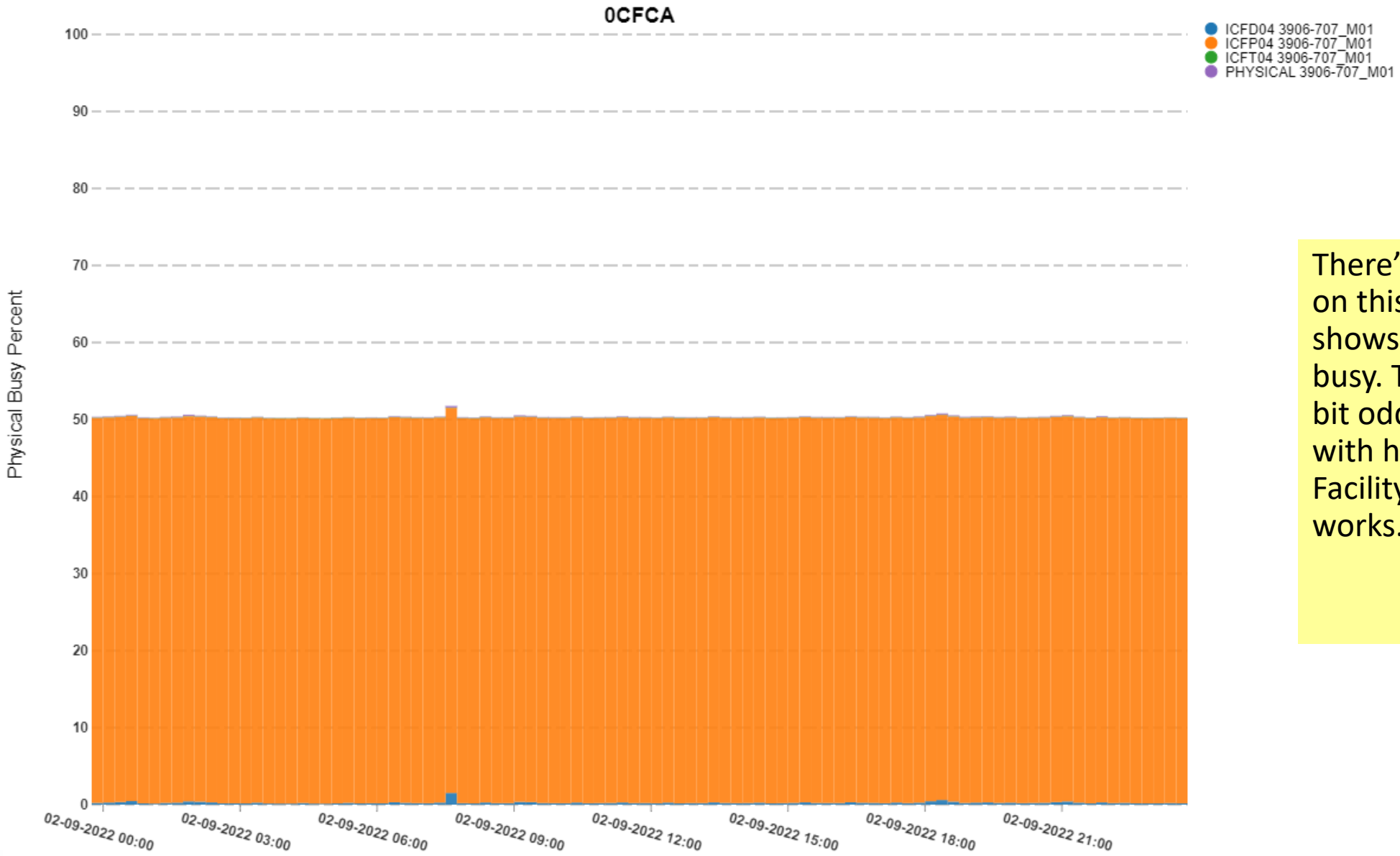
CEC Physical Machine zIIP Busy%



For completeness, here's the zIIP Busy for the machine. In this case the machine happens to have 8 zIIPs, making this utilization level relatively uninteresting.

If SMT is enabled, note "utilization" will be less than "busy".

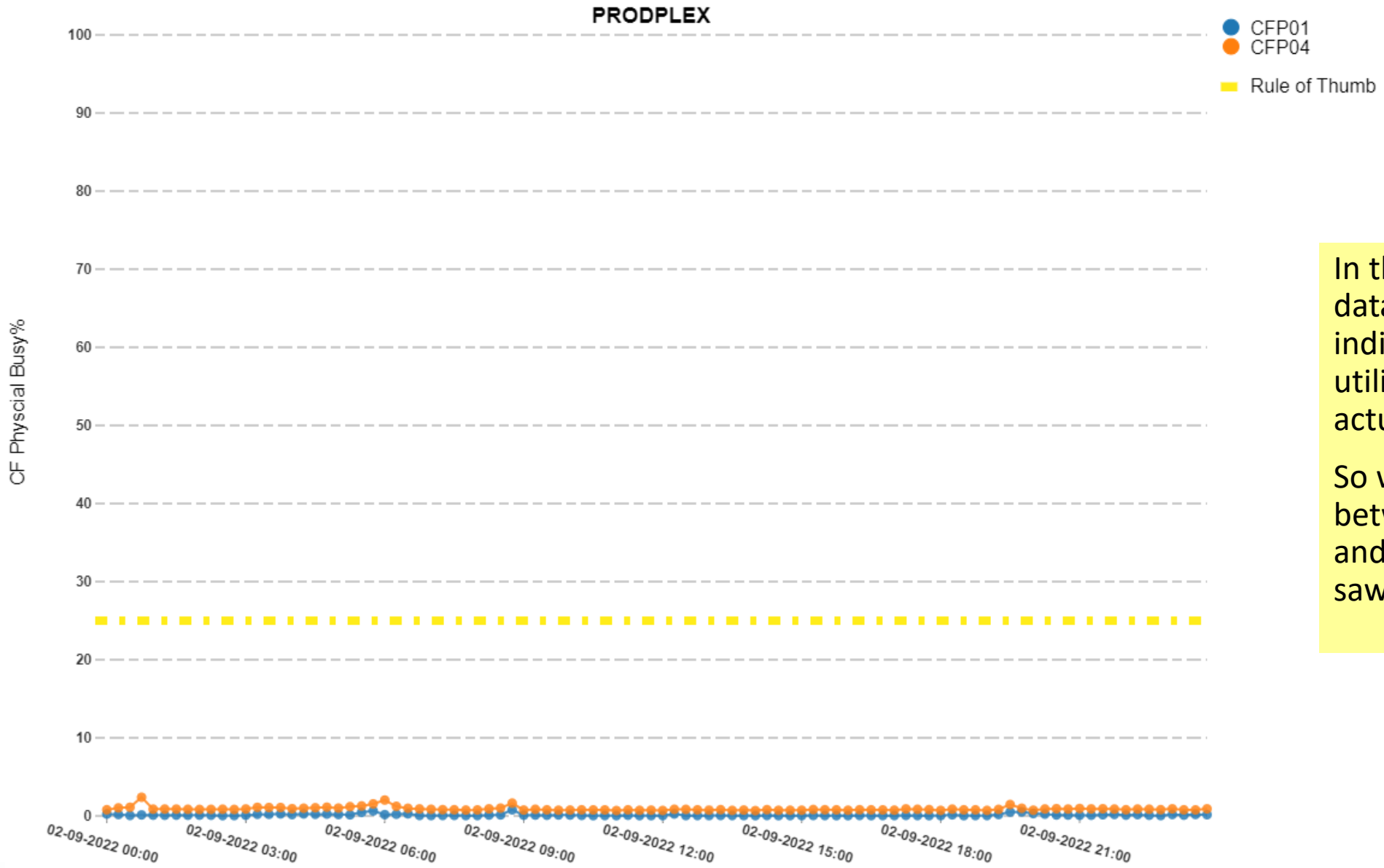
CEC Physical Machine ICF Busy%



There's two ICF engines on this box, and this shows that one is always busy. That might seem a bit odd, but it has to do with how the Coupling Facility Control Code works.

CF CPU - CF Processor Busy Utilization

As a Percentage of Physical Processor

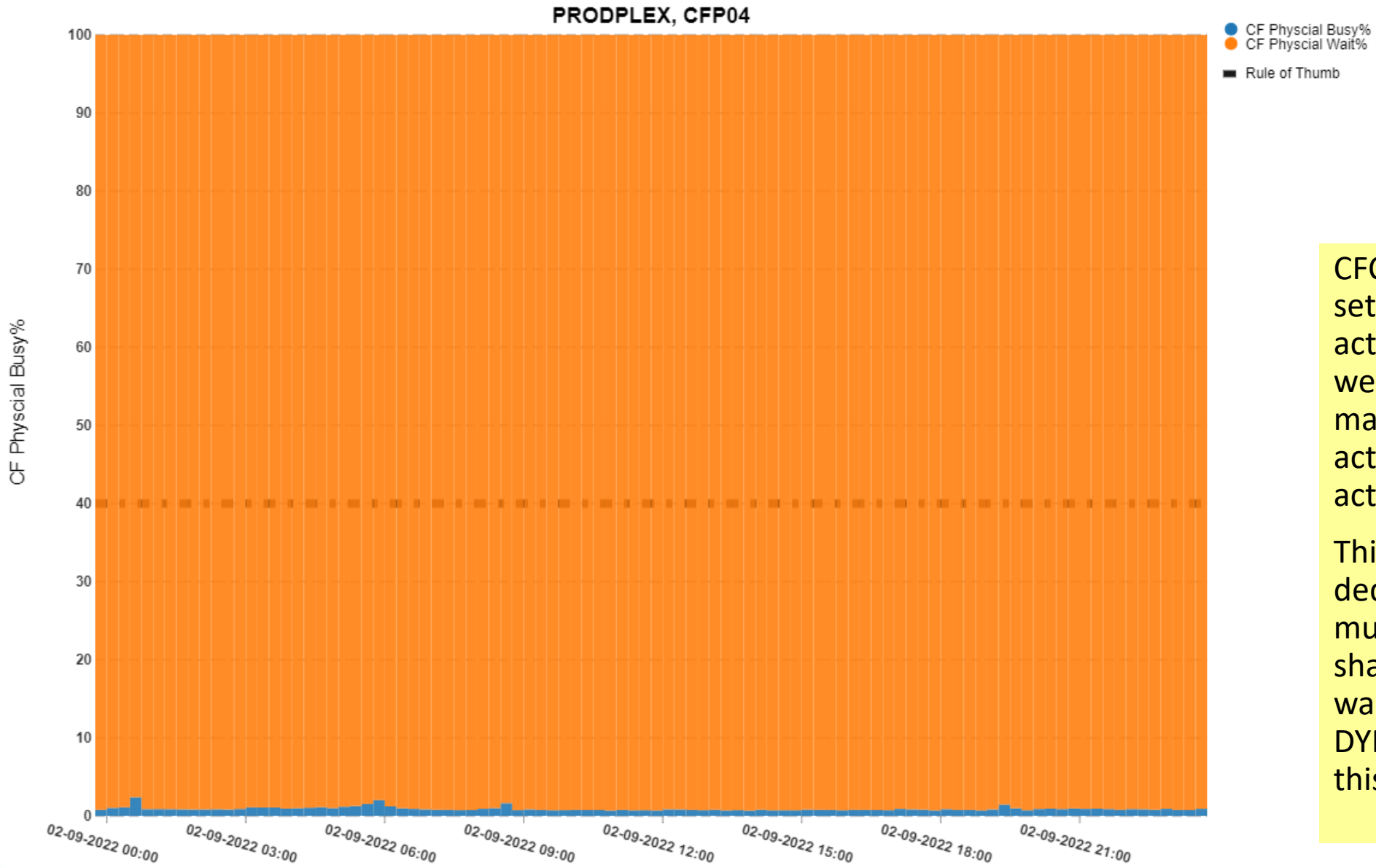


In this case the SMF 74 data gives us a better indication of how utilized the ICF engines actually are.

So what's the difference between this utilization and the busy number we saw on the prior report?

CF CPU - CF Processor (Busy and Wait) Utilization

As a Percentage of Physical Processor



CFCC (depending on settings) may run in an active wait loop. So here we see that the vast majority of its time is actually spent in that active wait loop.

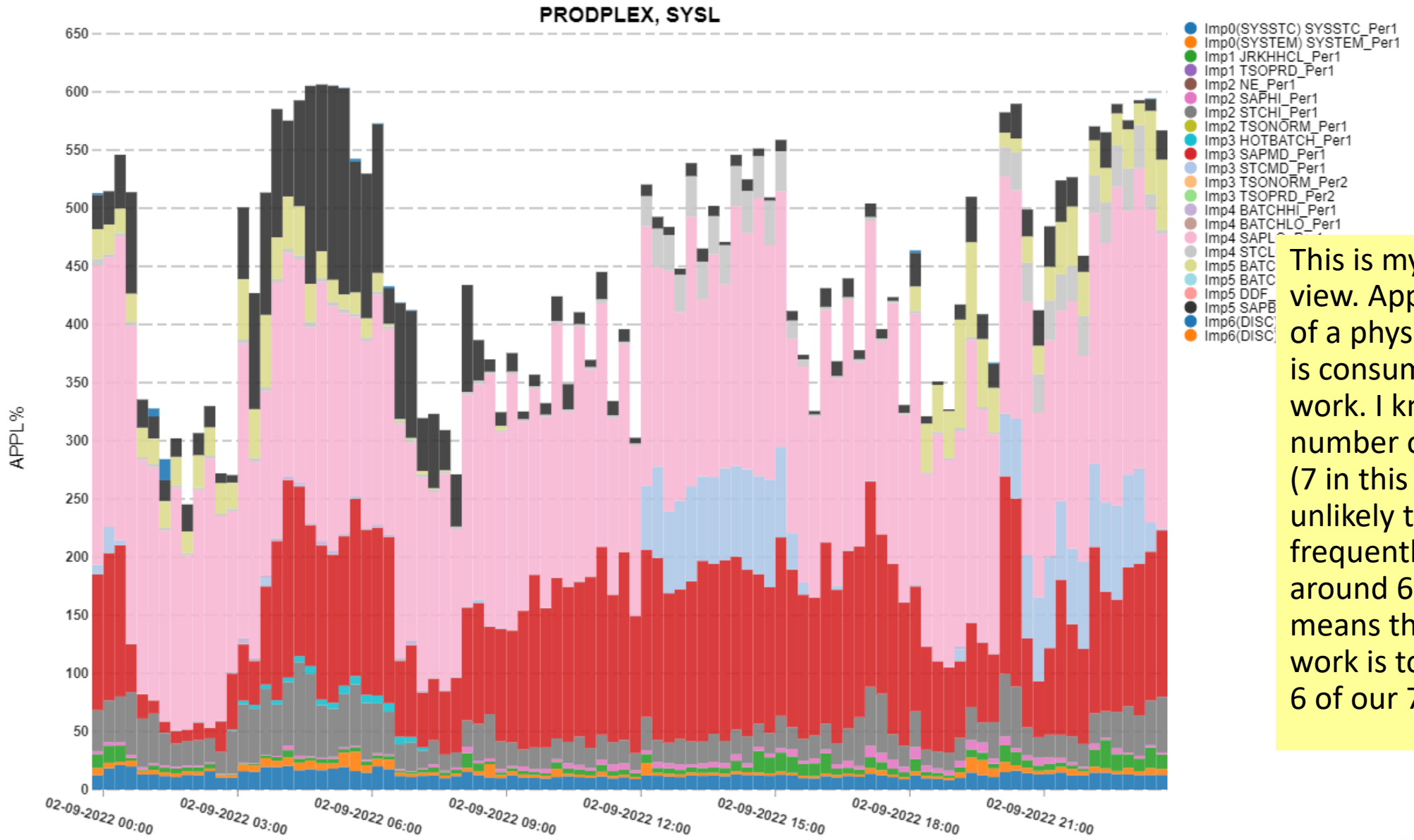
This is good if you have dedicated ICF engines. If multiple CF LPARs are sharing engines then you want to set DYNDISP=THIN to avoid this.

Going back to the GPs…

- Why did the machine seem to get busier in the past couple of weeks?
- The largest LPAR is PROD04 (aka SYSL)
- We need to look at the work running within the LPAR
  - Probably by service class, possibly by report class
- There are multiple measurements that reflect the same usage in different ways

WLM CPU - CP APPL% by Service Class

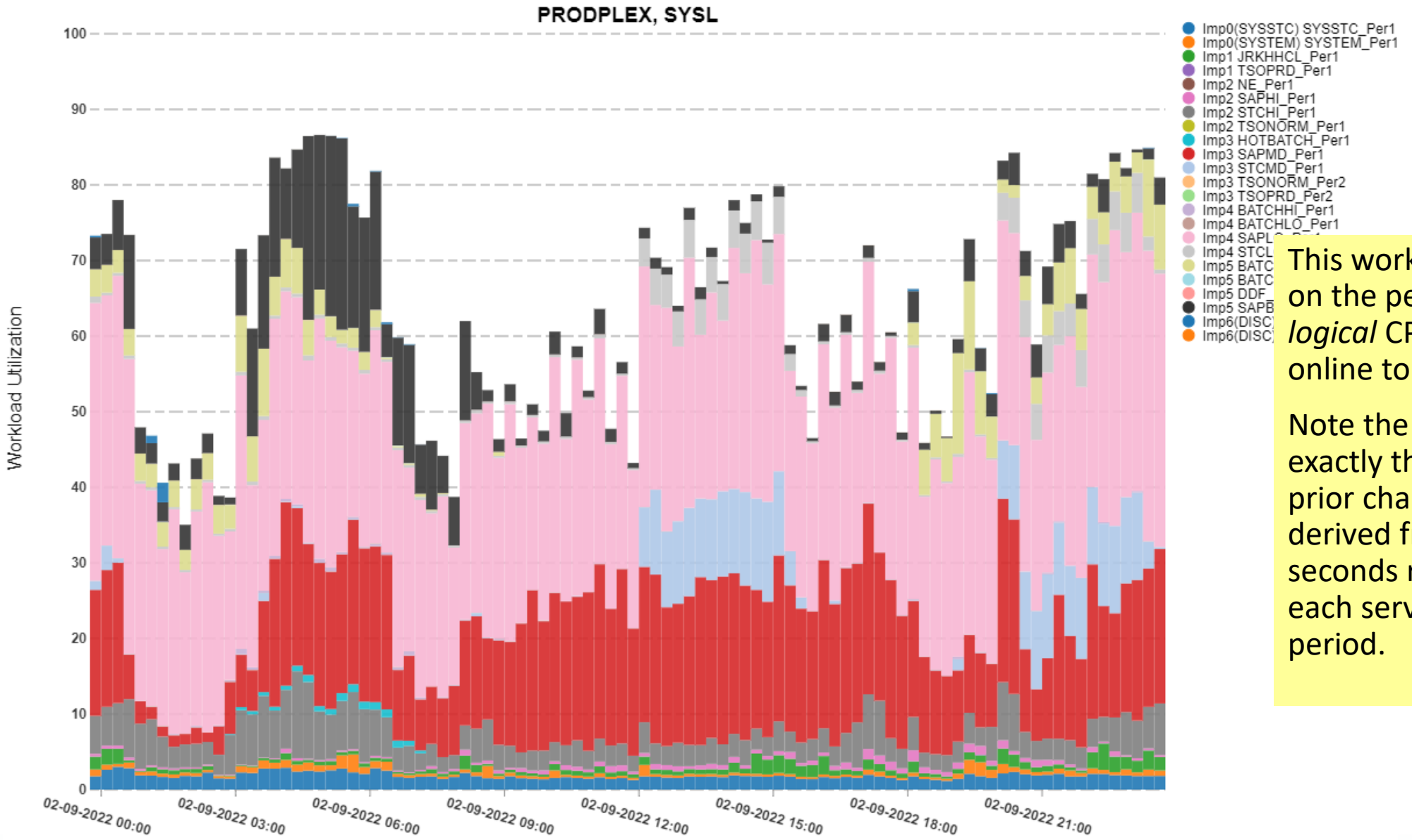
(CP + zAAP on CP + zIIP on CP)



This is my preferred view. Appl % is percent of a physical engine that is consumed by the work. I know the number of physical CPUs (7 in this case) and that's unlikely to change frequently. The peaks around 600 Appl% means that the LPAR's work is totalling to about 6 of our 7 CPUs.

WLM CPU - CP CPU Workload% by Service Class

(CP + zAAP on CP + zIIP on CP)



This workload % is based on the percent of the *logical* CPs that are online to the LPAR.

Note the profile is exactly the same as the prior chart: both are derived from the CPU seconds recorded for each service class period.

Logical vs. Physical Calculations

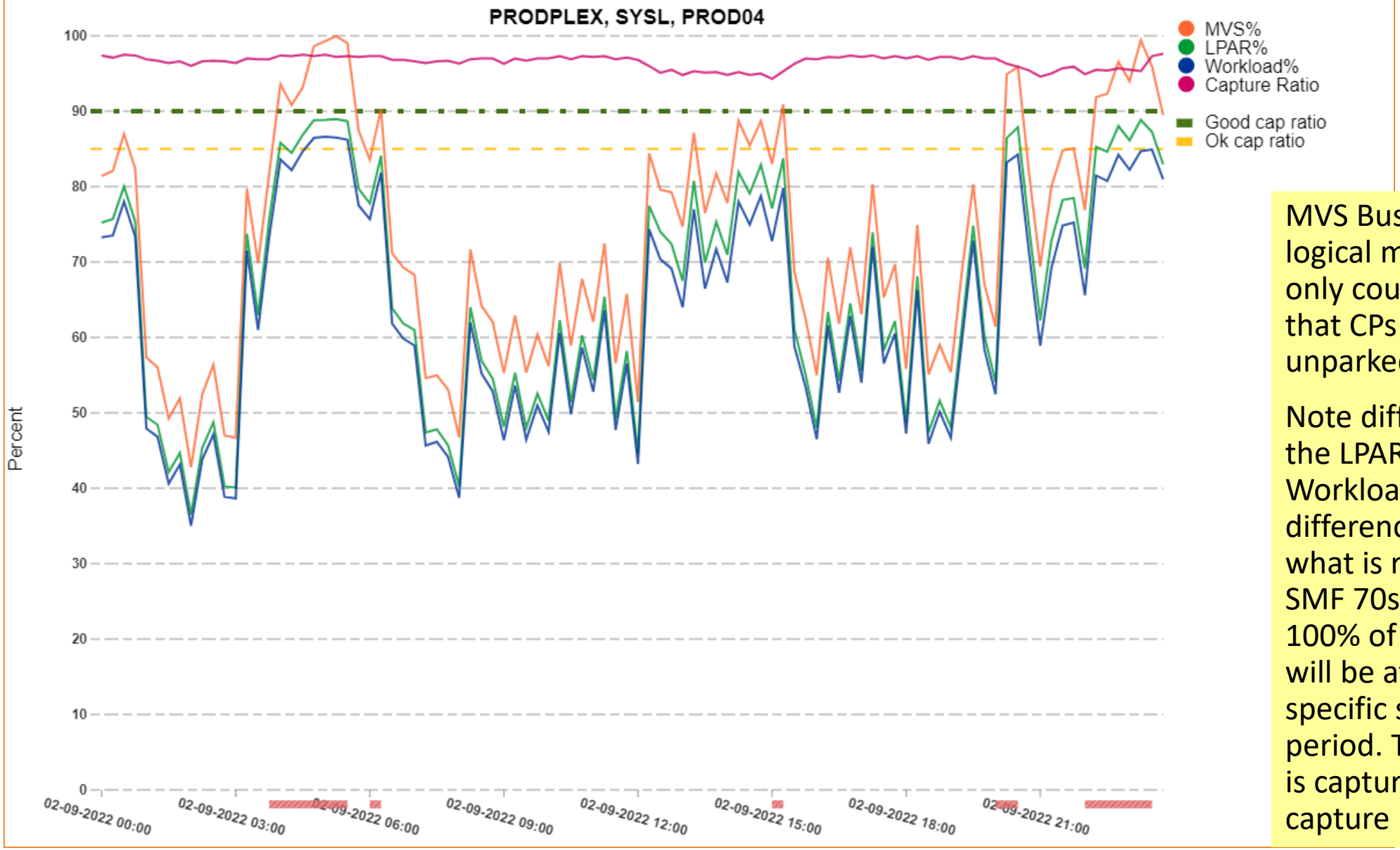
- 810 seconds of CPU time recorded for SC STCMD in 900 second interval
- 6 CPs online to the LPAR ("logical" CPs)
- 8 CPs characterized as GPs on the machine ("physical" CPs)

- APPL% = 810 / 900 = 90%
  - Percent of 1 CP's total capacity
- Logical Workload % = 810 / (900 * 6) = 15%
  - Percent of potential capacity of the online logical CPs
- Physical utilization % = 810 / (900 * 8) = 11.25%
  - Percent of the physical machine capacity

In the old days, when we tried not to define more logical CPs than absolutely necessary, monitoring logical CP utilization was more important.

Today, over-defining logical CPs by 1 or 2 and letting HiperDispatch manage them as vertical lows makes more sense and makes monitoring logical CP utilization less important.

LPAR, MVS, and Workload CP Busy% with Capture Ratio

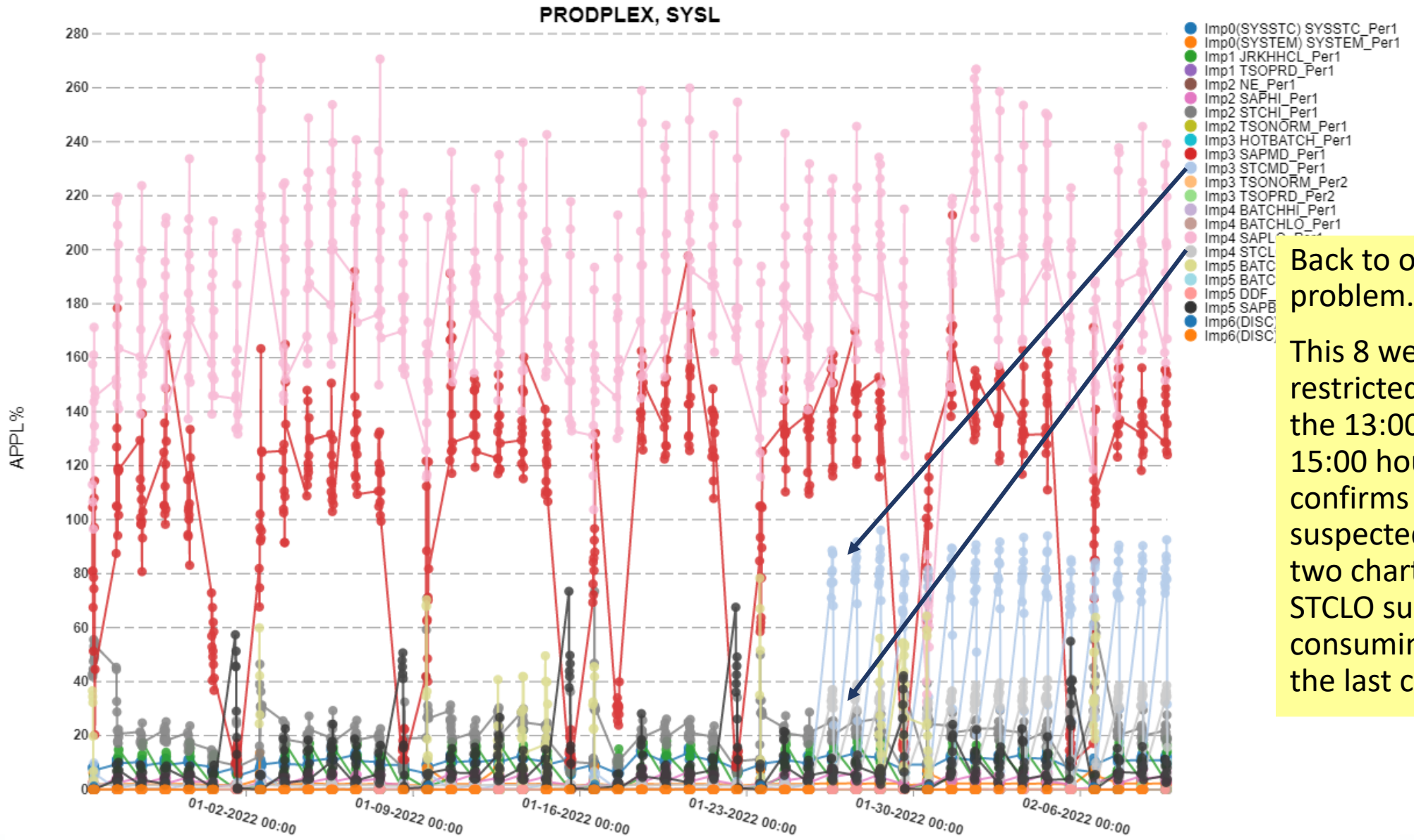


MVS Busy is another logical measurement but only counts the time that CPs were online and unparked.

Note difference between the LPAR% and Workload% is the difference between what is reported on the SMF 70s vs SMF 72s. Not 100% of the LPAR's time will be attributed to a specific service class period. The percent that is captured is the capture ratio.

WLM CPU - CP APPL% by Service Class

(CP + zAAP on CP + zIIP on CP)

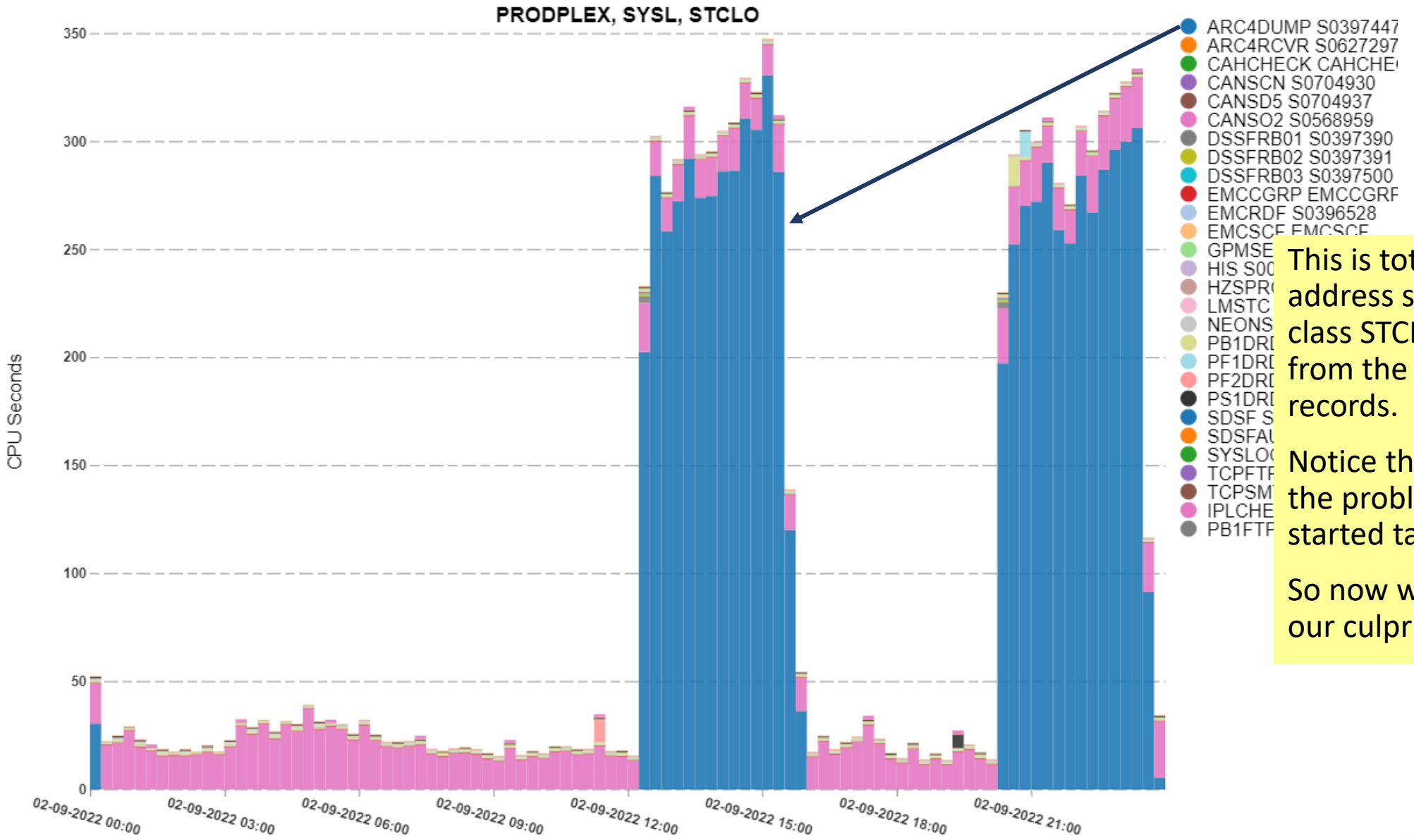


Back to our example problem…

This 8 week view was restricted to just show the 13:00, 14:00, and 15:00 hours. This confirms what we suspected from the prior two charts: STCMD and STCLO suddenly started consuming more CPU in the last couple of weeks.

Top Address Space CPU Time for Service Class

Period of Study



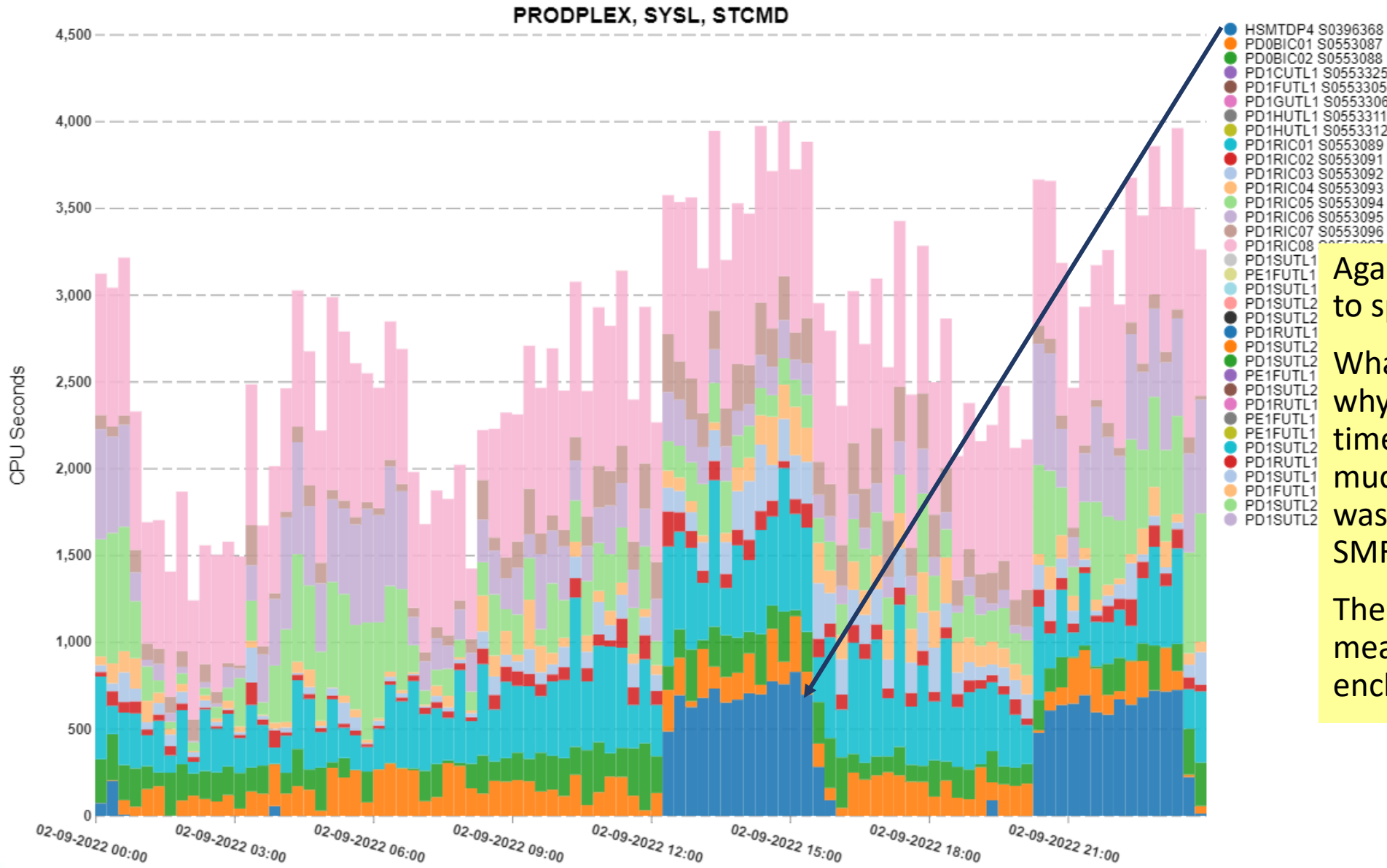
This is total CPU for address spaces in service class STCLO. This comes from the SMF 30 interval records.

Notice the big change in the problem hours for started task ARC4DUMP.

So now we know one of our culprits.

Top Address Space CPU Time for Service Class

Period of Study



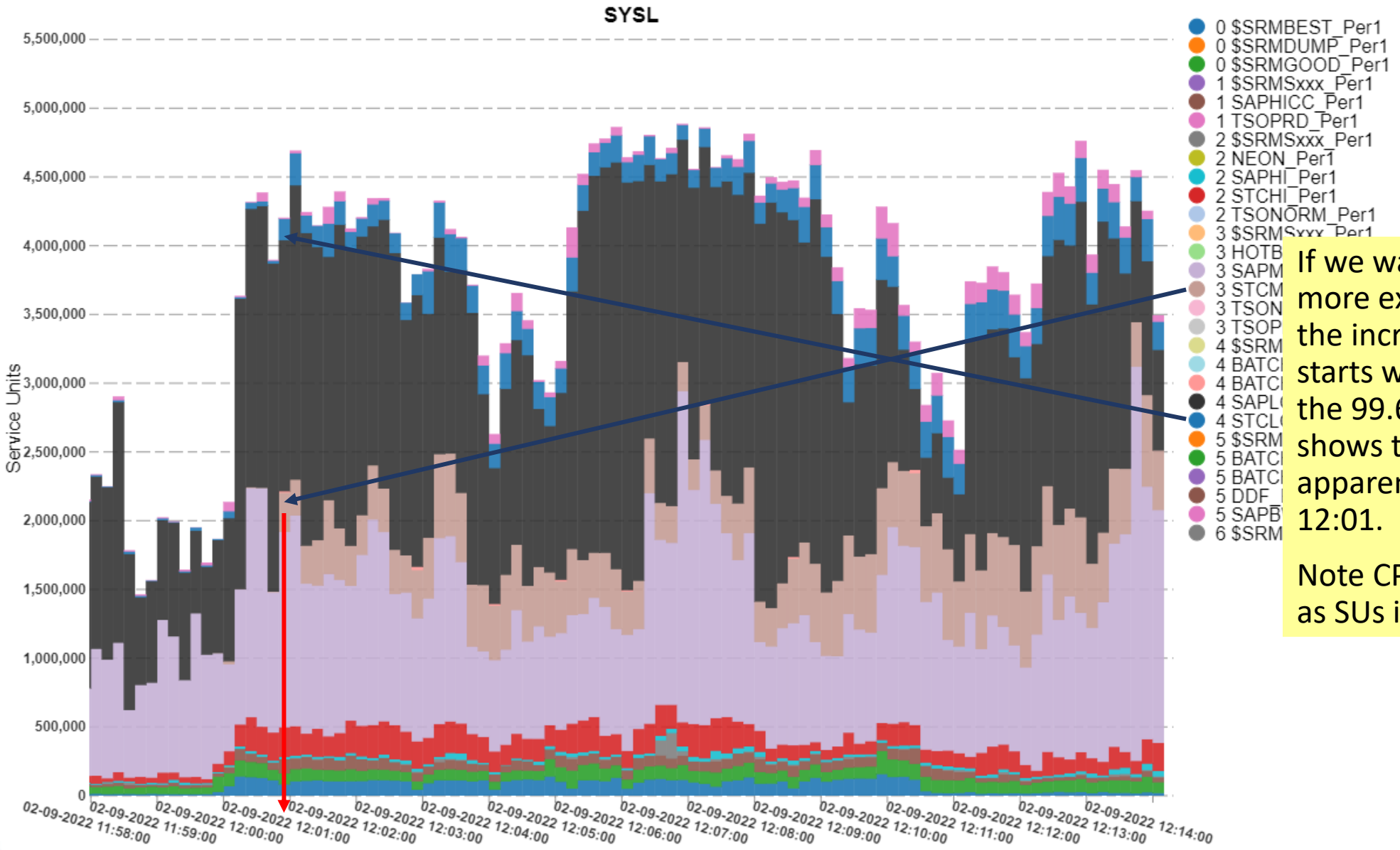
Again, the culprit is easy to spot.

What is interesting is why the SMF 30 CPU time seems to be so much higher than what was reported on the SMF 72s.

The answer to that means understanding enclaves.

CPU Accumulated by Service Class Period

From SMF 99.6



If we wanted to find a more exact time when the increased utilization starts we could look at the 99.6 data, which shows that those 2 SCs apparently get busy at 12:01.

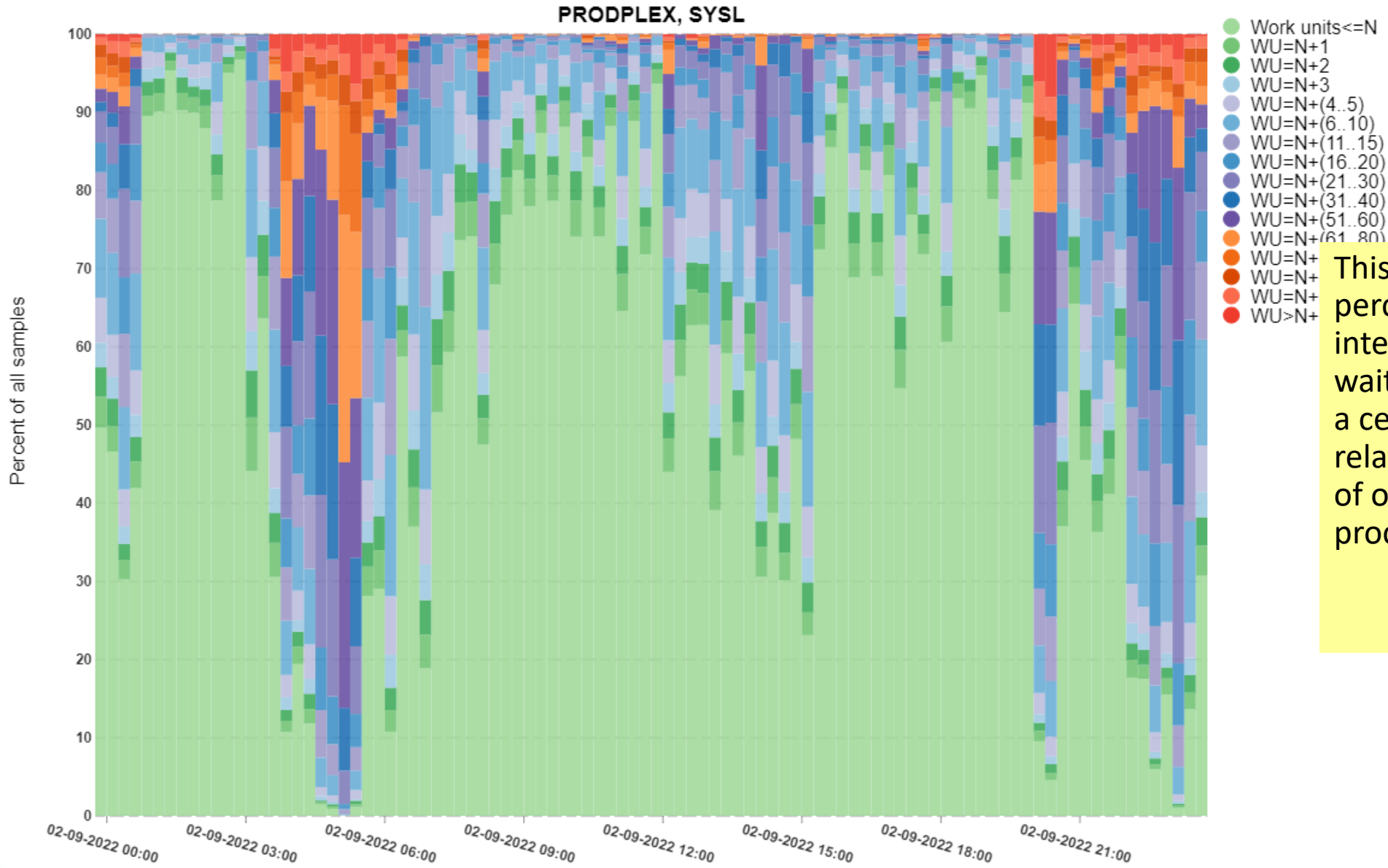
Note CPU time recorded as SUs in this record.

Does running at 100% hurt?

- Yes:
  - CPU time elongates as CPU utilization increases (due to contention)
- No:
  - WLM helps the important work get done even at 100% busy
  - (Assuming you have a mix of importances)
- Maybe:
  - Even when there is a mix of importances, there will be delays
  - Are the delays increasing, indicating that capacity is becoming more constrained?
  - Are our indications of latent demand increasing?

CPU Work Unit Distribution

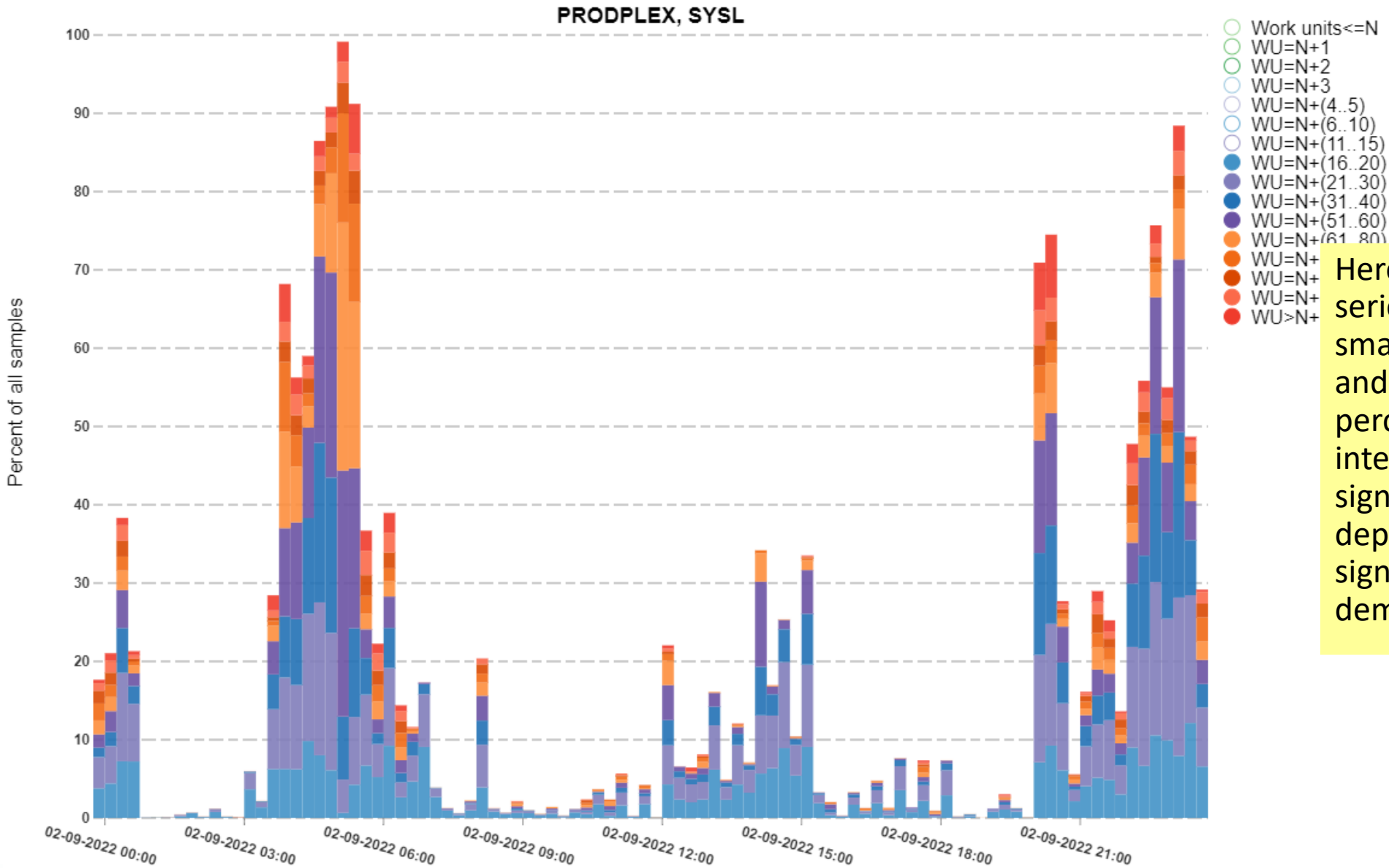
(N = Average of Unparked Engines Regardless of Engine Type)



This shows what percentage of the interval the running or waiting work units are at a certain number relative to the number of online, unparked processors.

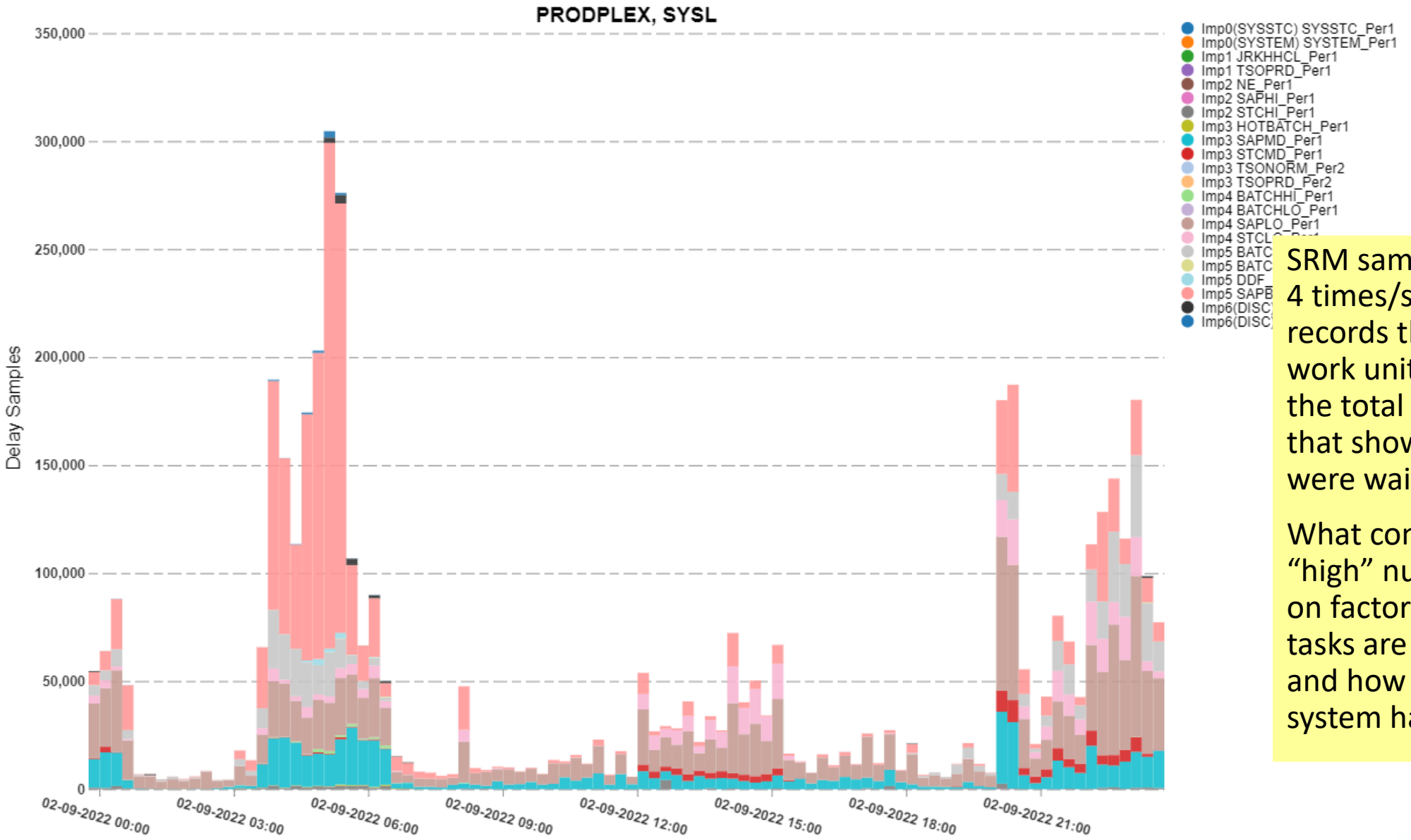
CPU Work Unit Distribution

(N = Average of Unparked Engines Regardless of Engine Type)



Here I've turned off the series for the relatively smaller queue depths and are left with the percentage of the interval that we had significant queue depths, indicating significant latent demand.

WLM CPU - CP CPU Delay Samples By Period

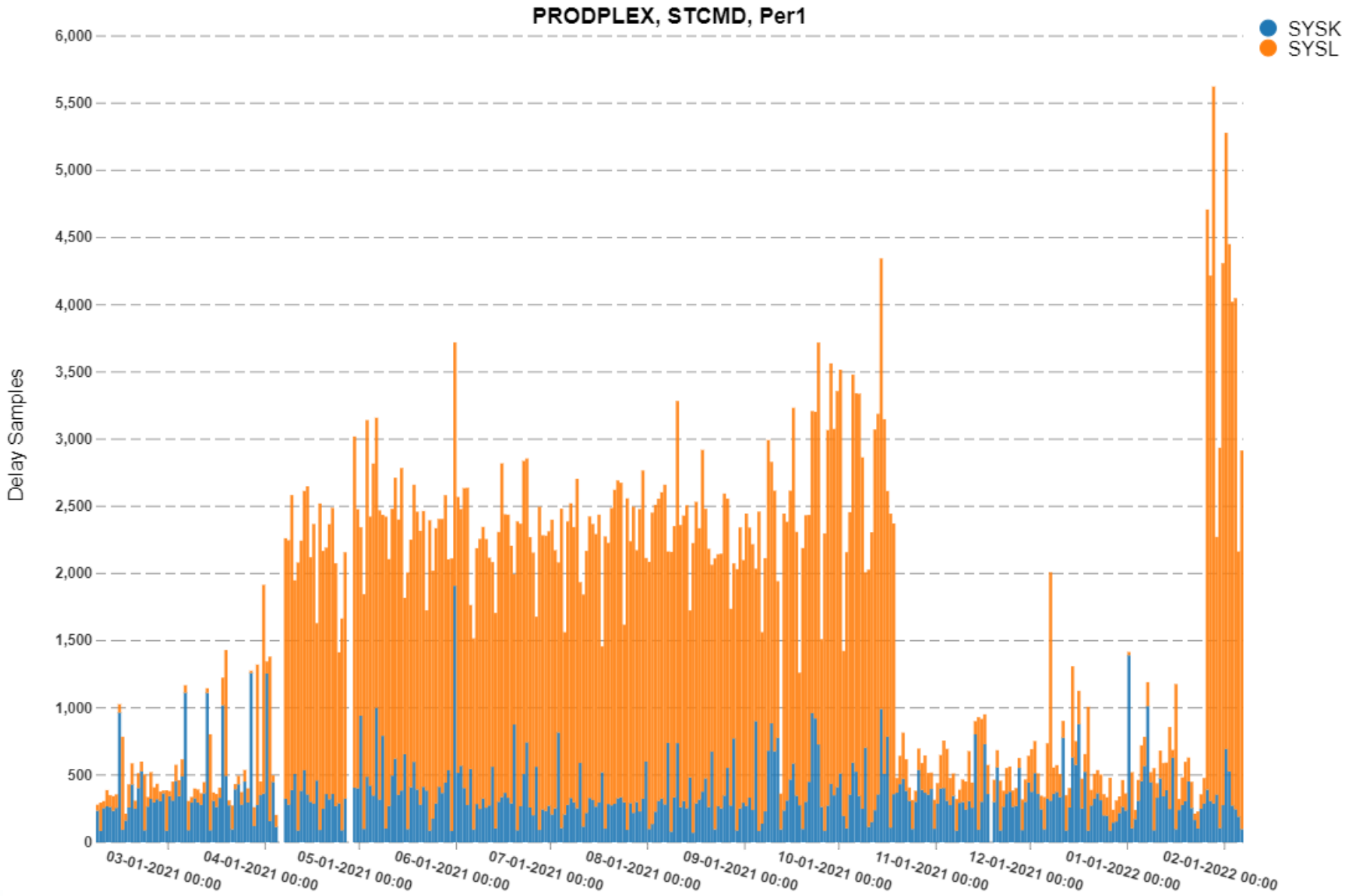


SRM samples the system 4 times/second and records the state of all work units. This shows the total of the samples that showed work units were waiting for CPU.

What constitutes a "high" number depends on factors like how many tasks are in the system and how many CPs the system has.

Daily 90th Percentile CPU Delay Samples

By System



This rolling year report shows us that STCMD has experienced more delays in the past couple of weeks, but that was after a period of a couple of months of relative quiet.

The reason for this pattern would require more investigation: was something fixed only to break again in the past couple of weeks? Is this somehow related to a business cycle?

Summary

What have we learned?

We've learned…

- CPU measurements are found in many SMF records
- CPU time is the basis for other CPU measurements
- MIPS, MSUs, and SU/sec all are measures of capacity and derive from the same artificial tests
  - MIPS & MSUs used for software pricing, SU/sec used internally by z/OS
- CPU time can be transformed to other more meaningful measures
  - Percent Busy, APPL %, MIPS, MSUs, etc.
- Percent Busy for zIIPs and ICFs may be only part of the story
- CPU delay samples can be useful to determine the relative amount of contention and whether that's increasing over time or not

Questions??

SHARE
EDUCATE • NETWORK • INFLUENCE

Complete your session evaluations for a chance at daily prizes!

To complete, visit
www.share.org/evaluation
and see your progress on the leaderboard!