Understanding SMF 98 Address Space Consumption 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

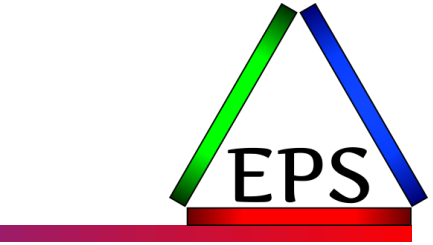
Abstract (why you're here!)

As we have previously discussed, the SMF 98 records are still a fairly new record, and there is still so much to learn about these high-frequency measurements. One set of measurements provided in the SMF 98 record are address space consumption measurements, which contain information about work unit dispatch times, dispatch counts, execution efficiency, and spin lock data for various priority buckets. What does all this mean? Well, we are not too sure yet, so attend this webinar with **Scott Chapman** to hear the results of our exploration of these measurements.

Agenda

- SMF 98 History
- Details on how the address spaces are selected
- Some thoughts on what this is good for
- Some example reports (of course)

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
    https://www.pivotor.com/content.html
    https://www.pivotor.com/webinar.html

z/OS Performance workshops available

During these workshops you will be analyzing your own data!

- WLM Performance and Re-evaluating Goals
  - May 12 – May 16, 2025 (4 days)
- Essential z/OS Performance Tuning
  - September 22-26, 2025 (4 days)
- Parallel Sysplex and z/OS Performance Tuning
  - October 21-22, 2025 (2 days)
- Also… please make sure you are signed up for our free monthly z/OS educational webinars! (email contact@epstrategies.com)

SMF 98 Introduction

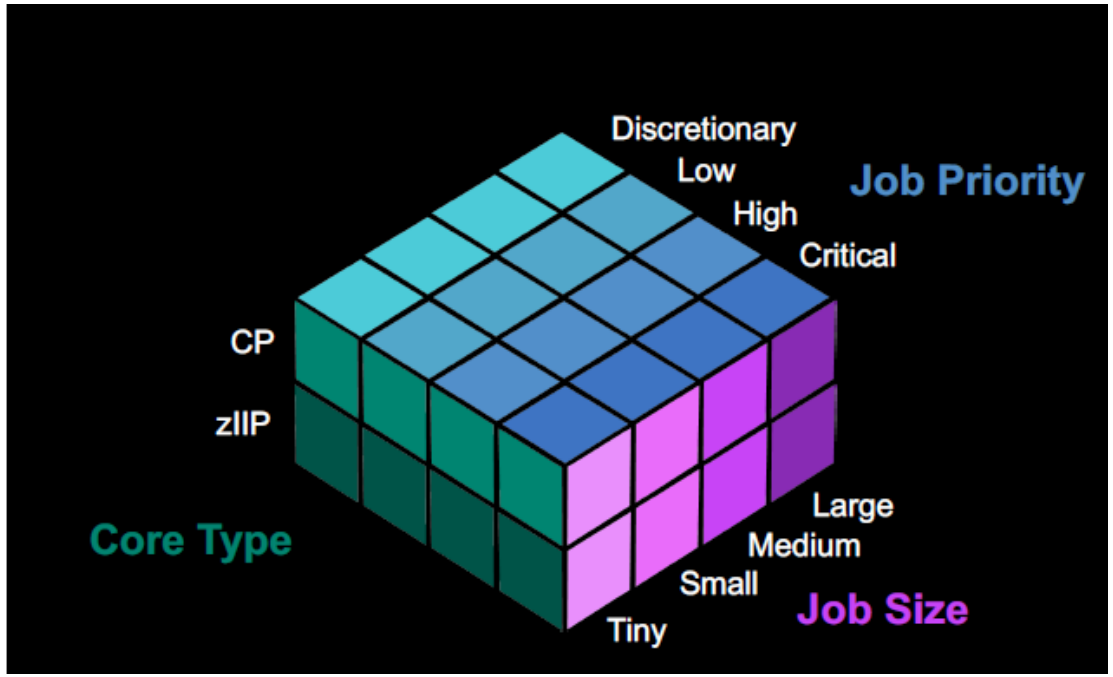
SMF 98 History

- The fine manual says SMF 98 is: “Workload interaction correlator and high frequency throughput statistics”
- Added for IBM z/OS Workload Interaction Correlator c. 2020
  - zWIC is a priced feature, but included with RMF or the Advanced Data Gatherer
    - Note Advanced Data Gatherer also included with RMF
  - z/OS Workload Interaction Navigator separately priced product for visualization
- High Frequency Throughput Statistics (subtype 1) are not dependent on WIC
  - Created by z/OS Supervisor
  - Contains interesting details about z/OS system and address spaces

Workload Interaction Navigator

The IBM solution

IBM z/OS Workload Interaction Navigator is a single interface to analyze Workload Interaction Correlator data. Correlates and recognizes multi-domain anomalous activity with cross-sectional views with exceptional job detail per cube.



From the IBM z/OS Workload Interaction Navigator Solution Brief
See: https://www.ibm.com/products/zos-workload-interaction-navigator

Remember this cube as it explains why the records are recorded the (nonintuitive) way they are.

SMF 98 Subtypes

Subtype	Record owner	Available with HFTS	Available with Correlator	Average record size per interval	Total average record data per day[1]
1	z/OS supervisor	Yes	Yes	32 KB	550 MB
3	z/OS I/O supervisor	No	Yes	2 KB	35 MB
4	z/OS I/O supervisor	No	Yes	2 KB	35 MB
5	DFSMS	No	Yes[2]	32 KB	550 MB
6	DFSMS	No	Yes[2]	32 KB	550 MB
7	DFSMS	No	Yes[2]	32 KB	550 MB
8	DFSMS	No	Yes[2]	32 KB	550 MB
1024	CICS	No	Yes	2 KB	35 MB
1025	IMS	No	Yes	2 KB	35 MB

May be slightly high from what we've seen

From z/OS SMF Manual

- To enable subtype 1 need HFTSINTVL in SMFPRMxx
  - No additional cost
- To enable others, need to enable WIC in IFAPRDxx
  - Also need to add parameter WIC in SMFPRMxx
  - Potential for extra cost if you don't license RMF/ADG

HFTS Interval

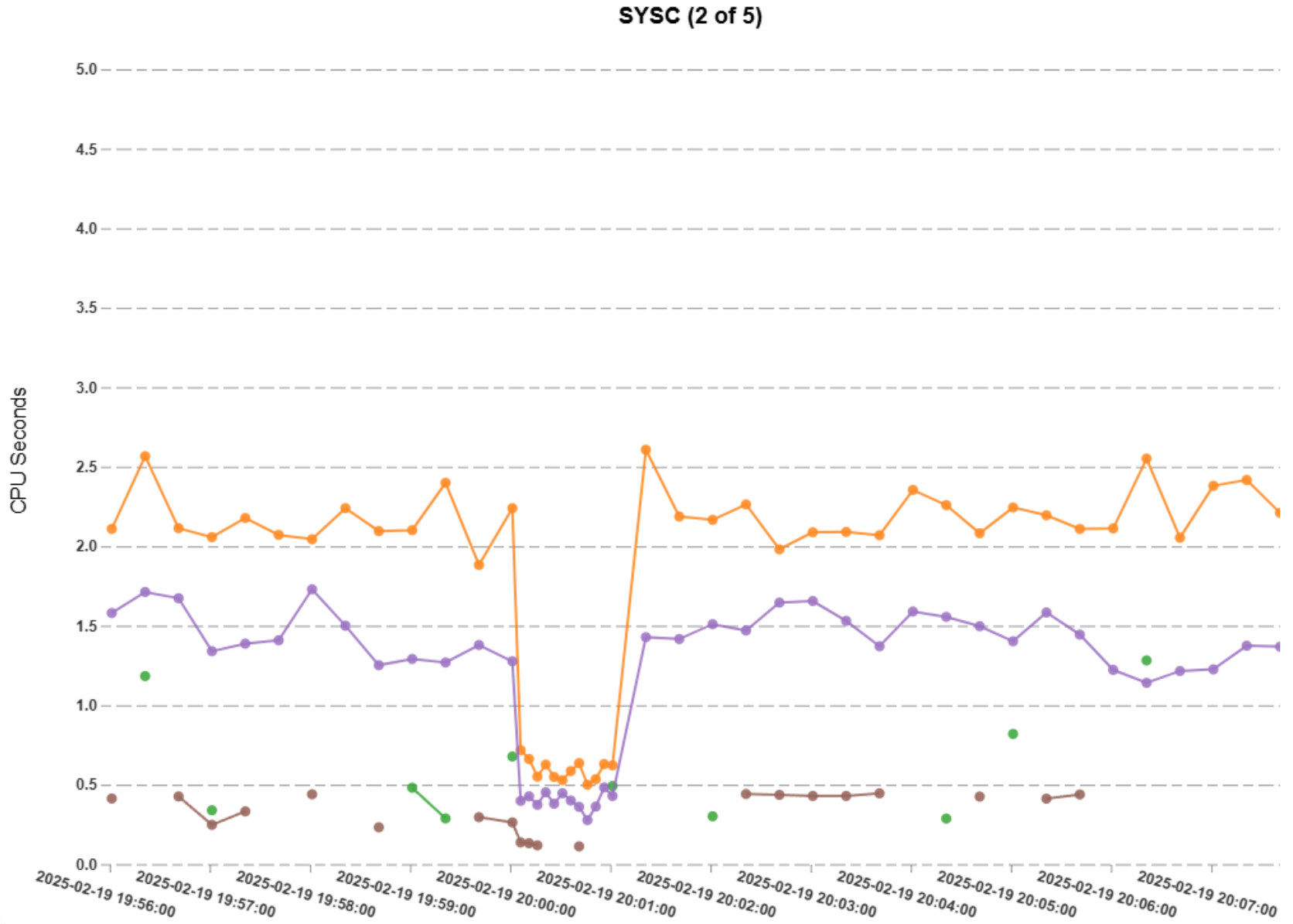

```
In SMFPRMxx:

HFTSINTVL(5)
```

- Default is NOHFTSINTVL, so have to set it to get records
- Can set it to 5, 10, 15, 20, 30, or 60 seconds
- If not set to 5 seconds:
  - an interval of 5 seconds will be used anyways for minutes 0, 15, 30, and 45
- IBM Recommendation is to make it 5 seconds
- Our recommendation is now 5 seconds too
  - The information can be useful and the cost of a few hundred MB is negligible
  - Larger intervals can make some reporting more confusing
- If WIC is specified in SMFPRMxx, then HFTSINTVL specification is ignored and 5 seconds is used

Address Space CPU Consumption

When Recorded on HFTS Interval



Note how the dots are closer together at the top of the hour.

That drop in in CPU used by TCP/IP and XCFAS by about 75% is due to the short intervals.

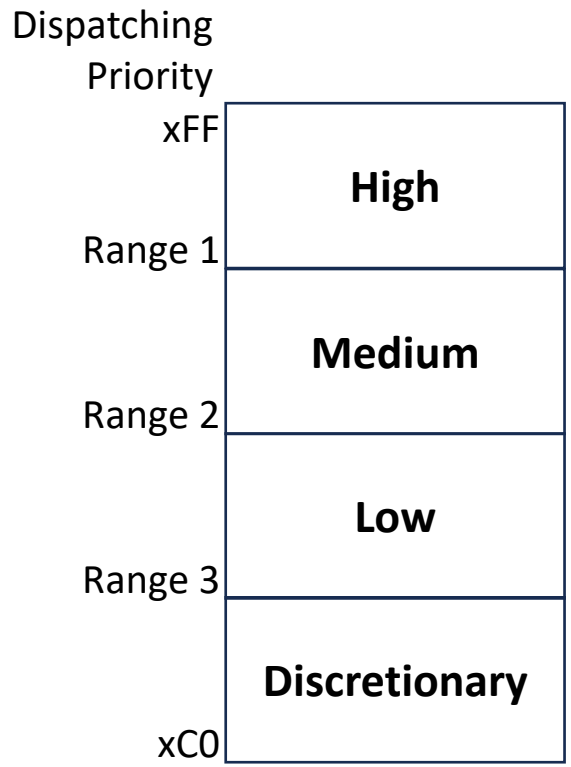
Also note VTAM and *MASTER* probably weren't idle when not shown… they just weren't recorded.

Address Space Recording

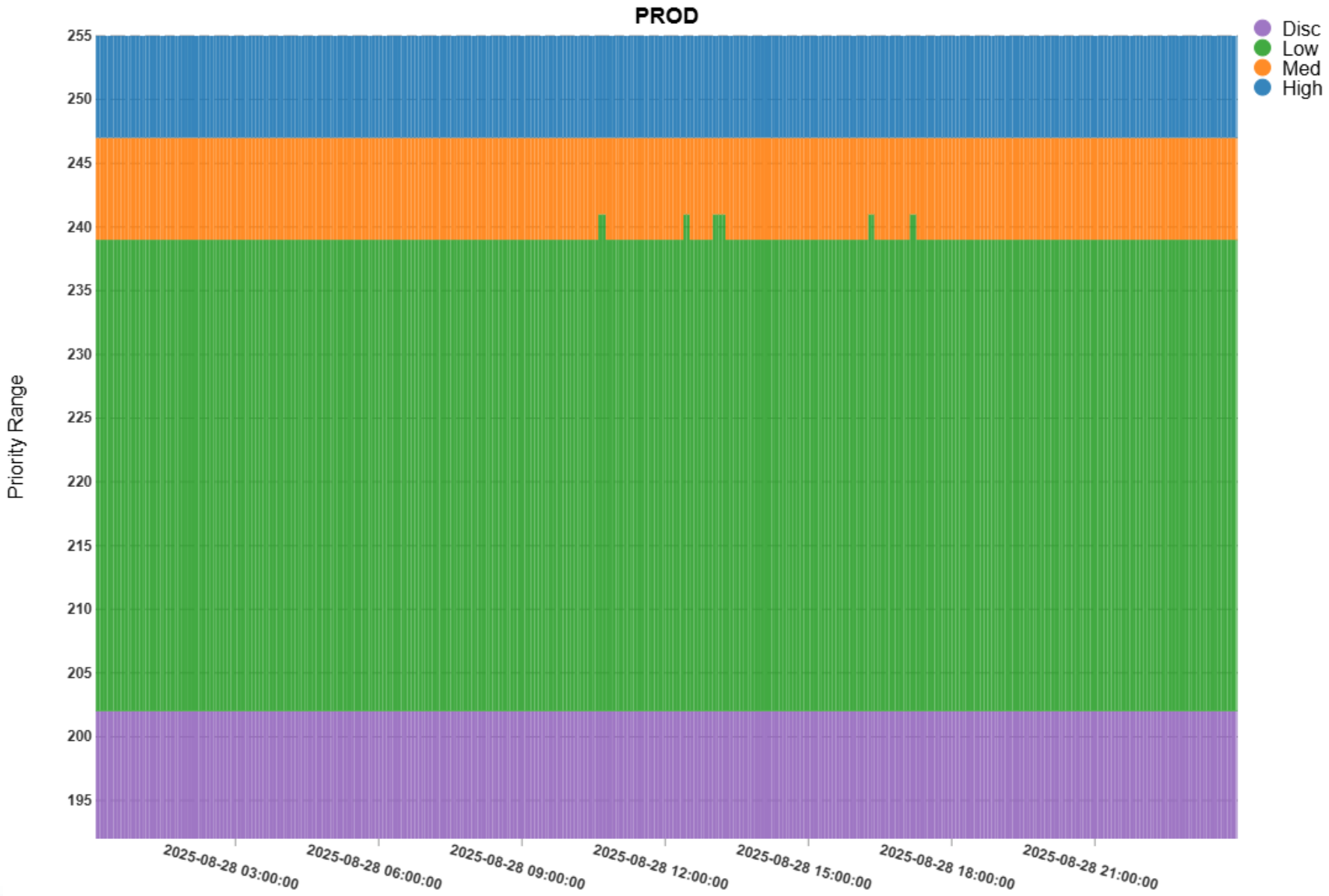
- Which address spaces get recorded is… confusing
- The top address space is recorded across 3-4 dimensions
  - Processor class (CP or zIIP)
  - Priority Bucket
  - Sub-bucket
  - Work Unit Type (for consumption, not efficiency or spin lock detail)
- The final 3 all also record which is the top address space across all buckets, sub-buckets, or work unit types
- In any given interval an address space may be referenced multiple times
  - E.G. as the top CP consumer in PB “high” for multiple work unit types
  - Absolute top will of course show up in both its relevant dimension and under “all”
- Some combinations may have no entries
  - E.G. maybe work in priority bucket 1 is only in sub-buckets 1, 2, and 4

Priority buckets

- Priority buckets are dispatching priority ranges
  - Specified as 3 range values
  - High = xFF to range(1)
  - Medium = range(1)-1 to range(2)
  - Low = range2-1 to range(3)
- Ranges do change sometimes
  - Presumably due to WLM changing dispatching priorities
  - But seems to be rare, so not sure what is really driving it
- Ensures that some address spaces across all priorities are captured
- Special priority bucket of “All” that crosses all priorities
  - Apparently to make the Navigator’s reporting job easier?



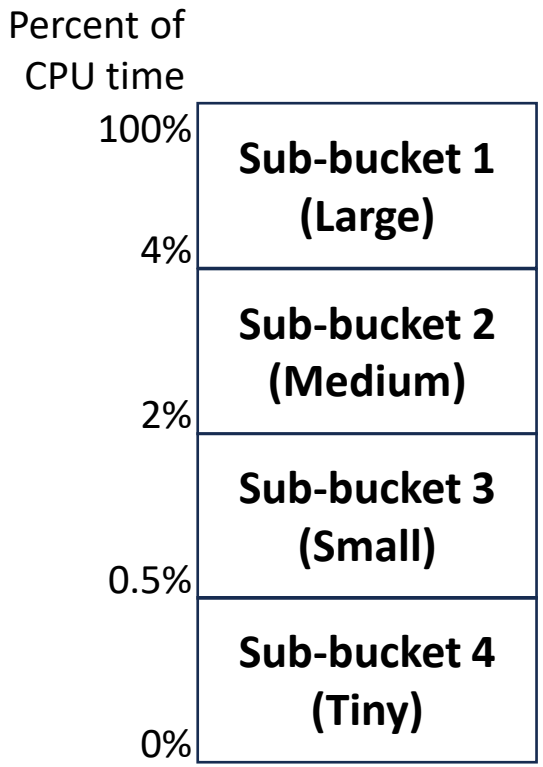
HFTS Priority Bucket Ranges



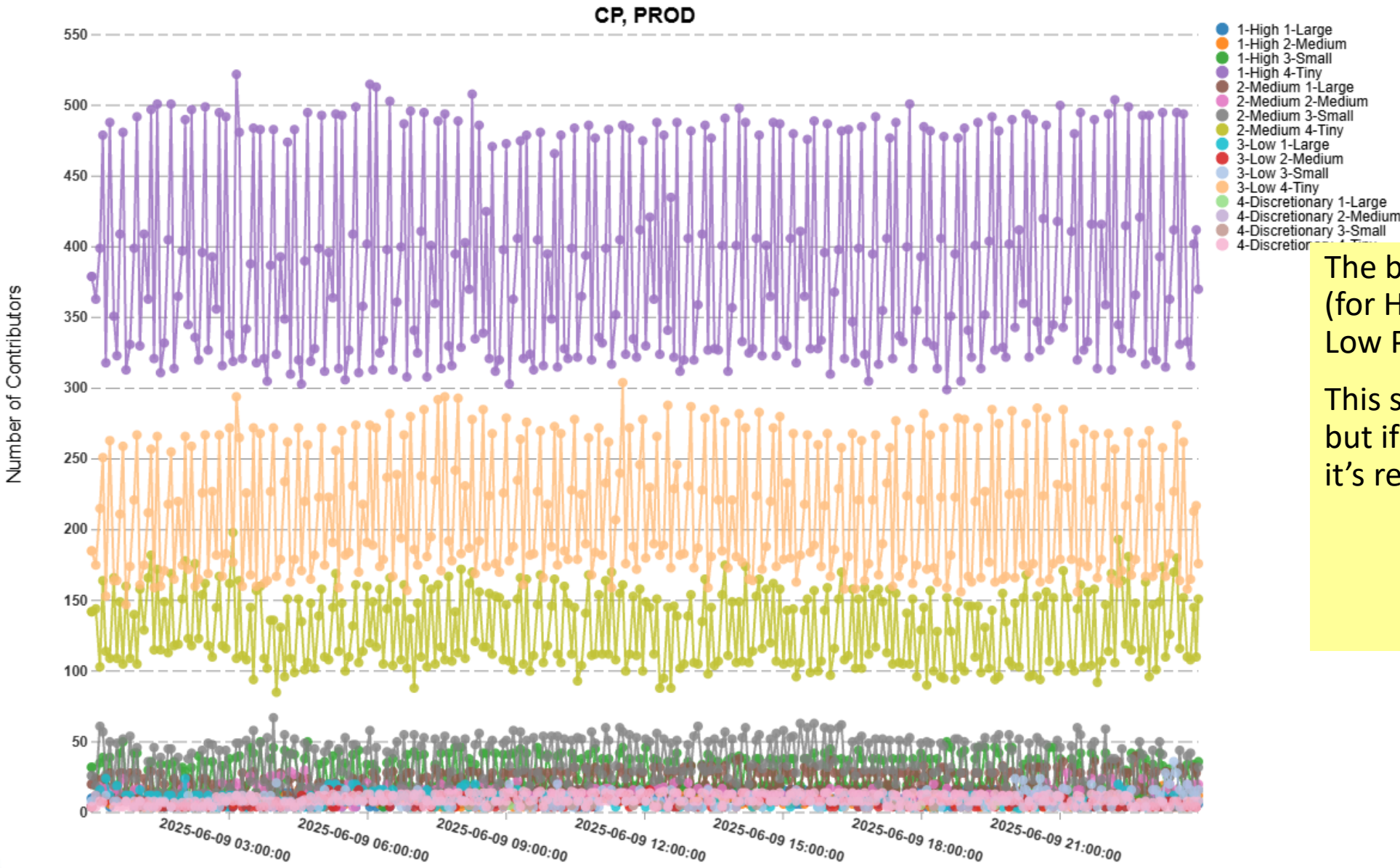
Note minor variation between the Medium and Low priority ranges

Sub-buckets

- "A sub-bucket is a collection of address spaces that consume similar CPU times relative to the total CPU time from the processor class."
- Specified by ranges of percentage of total CPU consumed within the processor class
- Can (theoretically) change, but I haven't seen that
  - 4%
  - 2%
  - 0.5%
- These ranges seem somewhat surprising
  - "Large" is "using more than 4% of the total"
- But works fairly well
  - Lots of work units in "tiny", remaining sub-buckets have somewhat similar contributor numbers
- Additional sub-bucket representing the entire range as well



HFTS Priority & Sub- Bucket Contributor Count



The big bands are all tiny (for High, Medium, and Low Priority Buckets)

This seems surprising, but if you think about it, it's really not.

Work unit type

- For the consumption data, an additional dimension is used to find the address spaces based on the top work unit types:
  - All tasks/SRBs
  - Non-enclave task
  - Enclave task and SRB
  - Non-enclave, pre-emptible SRB
  - Non-pre-emptible SRB
  - “Reserved for IBM Use”
    - Data does show up for this selection and does select potentially different address spaces

Design implications

- Only a subset of address spaces selected per interval
  - Usually less than a few dozen out of hundreds on most systems
- Some address spaces qualify across more than one dimension
  - “All” dimensions always contains a duplicate
- Address spaces are not (necessarily) continuously recorded
  - Lack of data about an address space just means it wasn’t top in one of combos
- I wish they would have:
  - Included service class period as a dimension
  - Included more than just the top 1 address space

So what is it all good for?

Consumption

- We can see Service Class Period CPU consumption on a 10 second basis with SMF 99.6, the 98.1s potentially let us see Address Space consumption on a 5 second basis (for some, some of the time).

Efficiency

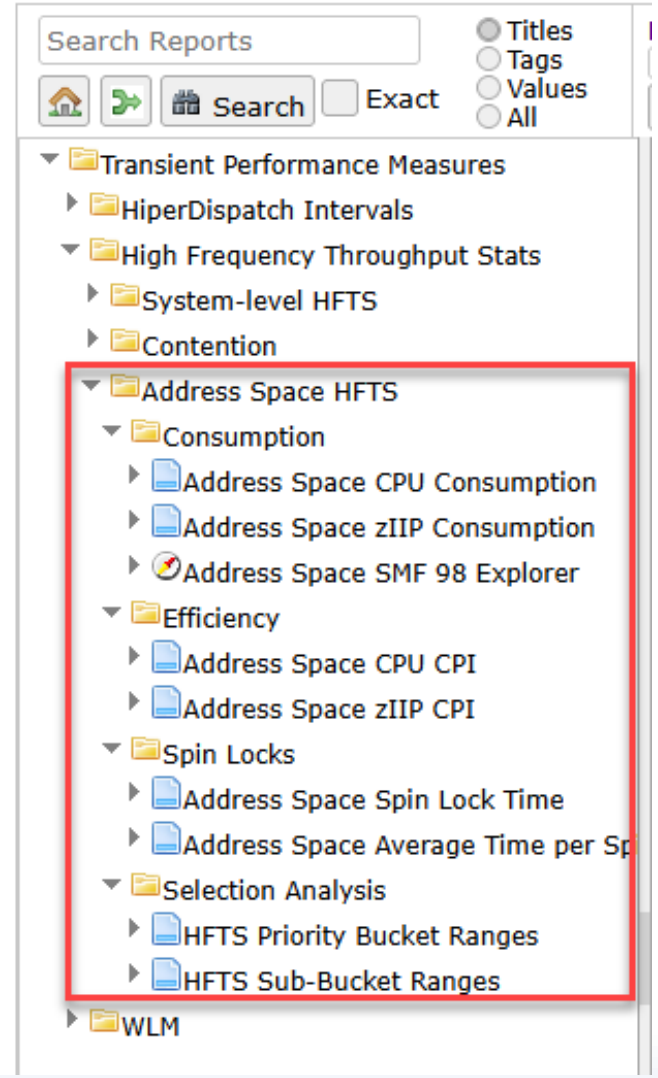
- We occasionally see systems with periodic SIIS issues indicating possibly some very old code being run. But that's hard to track down, even with the SMF 30 instruction counts. This might be another way of looking at it that could (maybe) be useful.

Spin Locks

- If you have an indication of a spin lock problem this may help identify where
- An increase in the average time per spin lock might also be interesting (maybe)

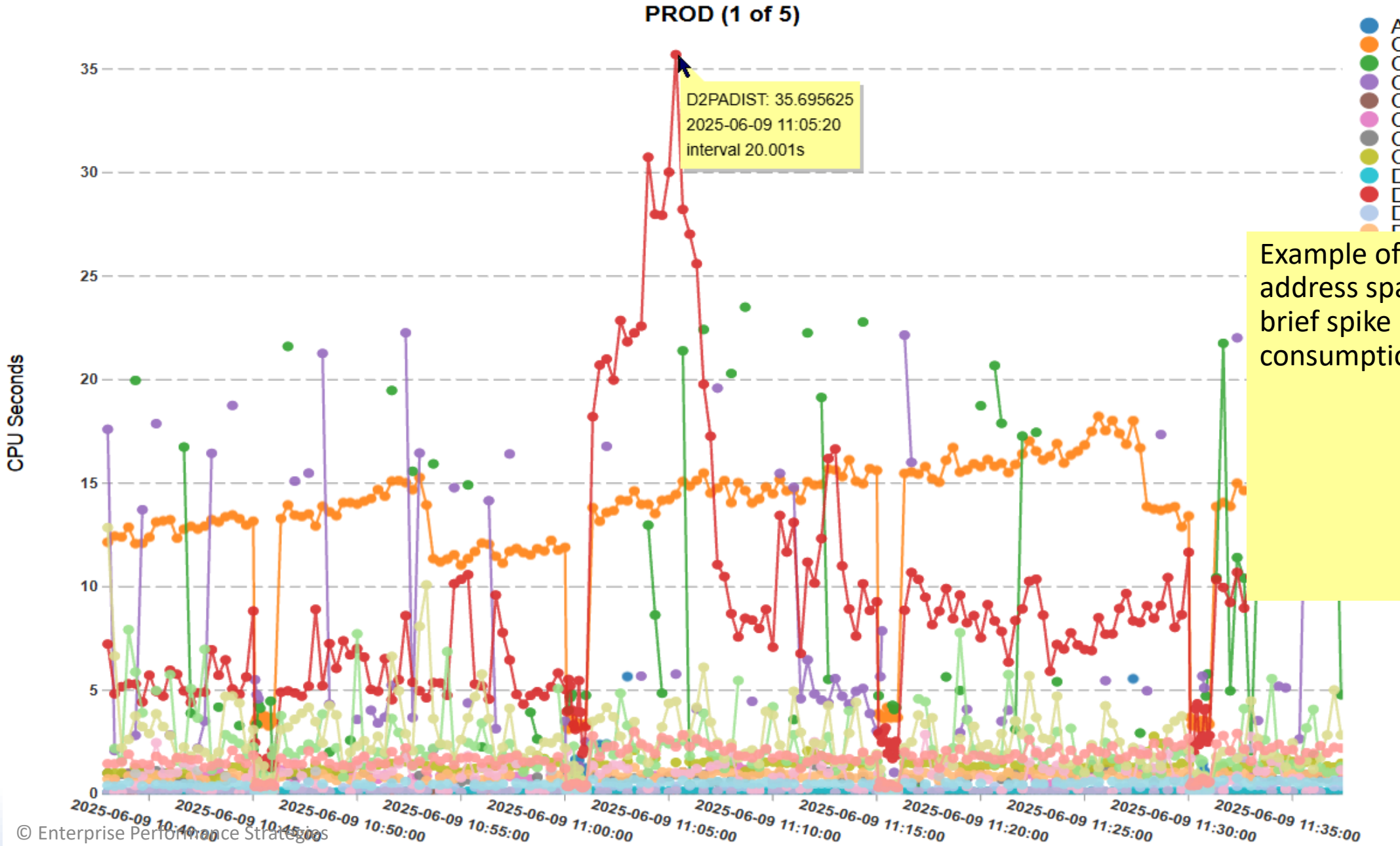
For Pivotor customers…

- See the Transient Performance Reportset
- This isn't quite everything we could report on, but it is probably what you might find interesting when looking into a problem
- Let us know if you have some other ideas



Address Space CPU Consumption

When Recorded on HFTS Interval

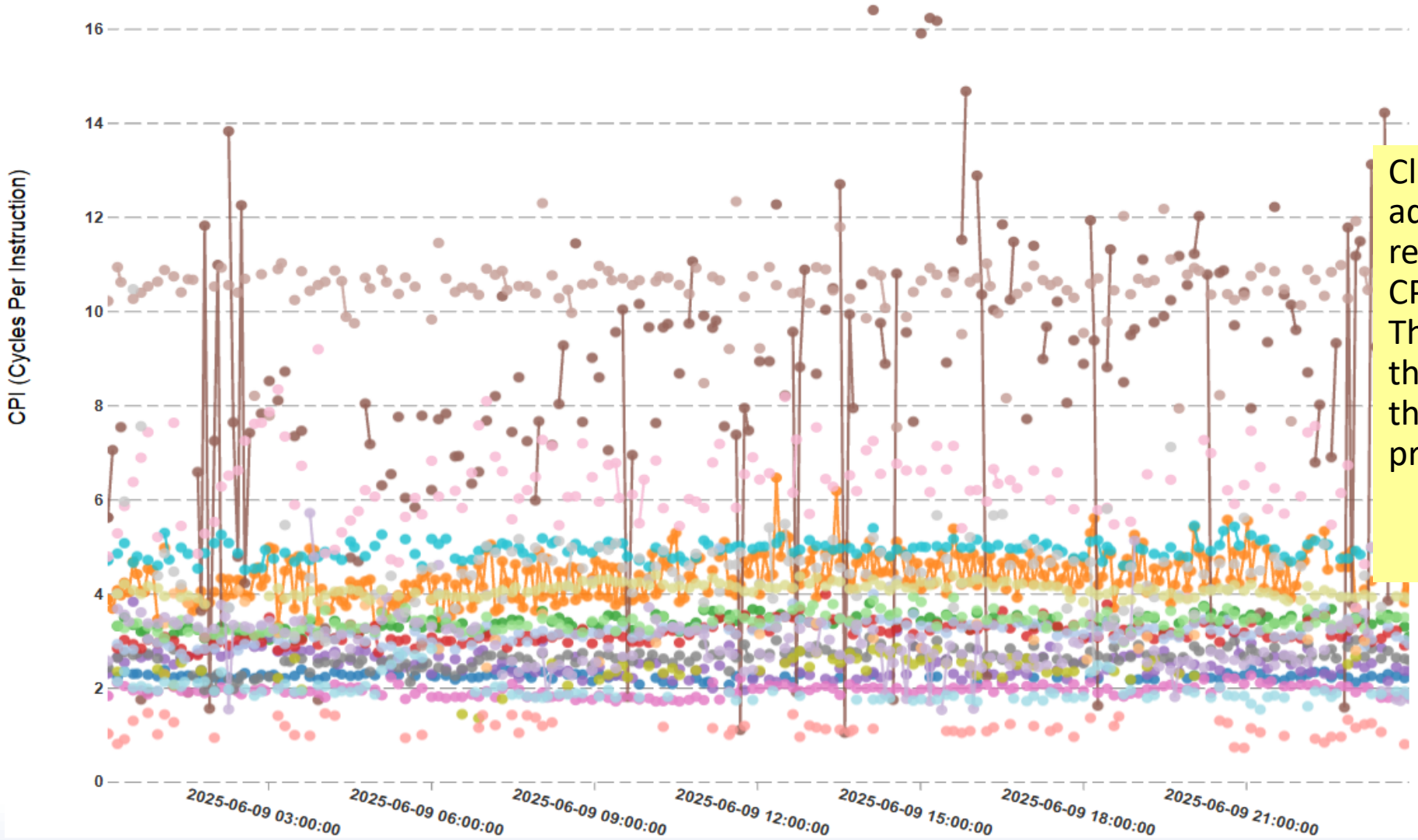


Example of how one address space had a brief spike in CPU consumption

Address Space CPU CPI

When Recorded on HFTS Interval

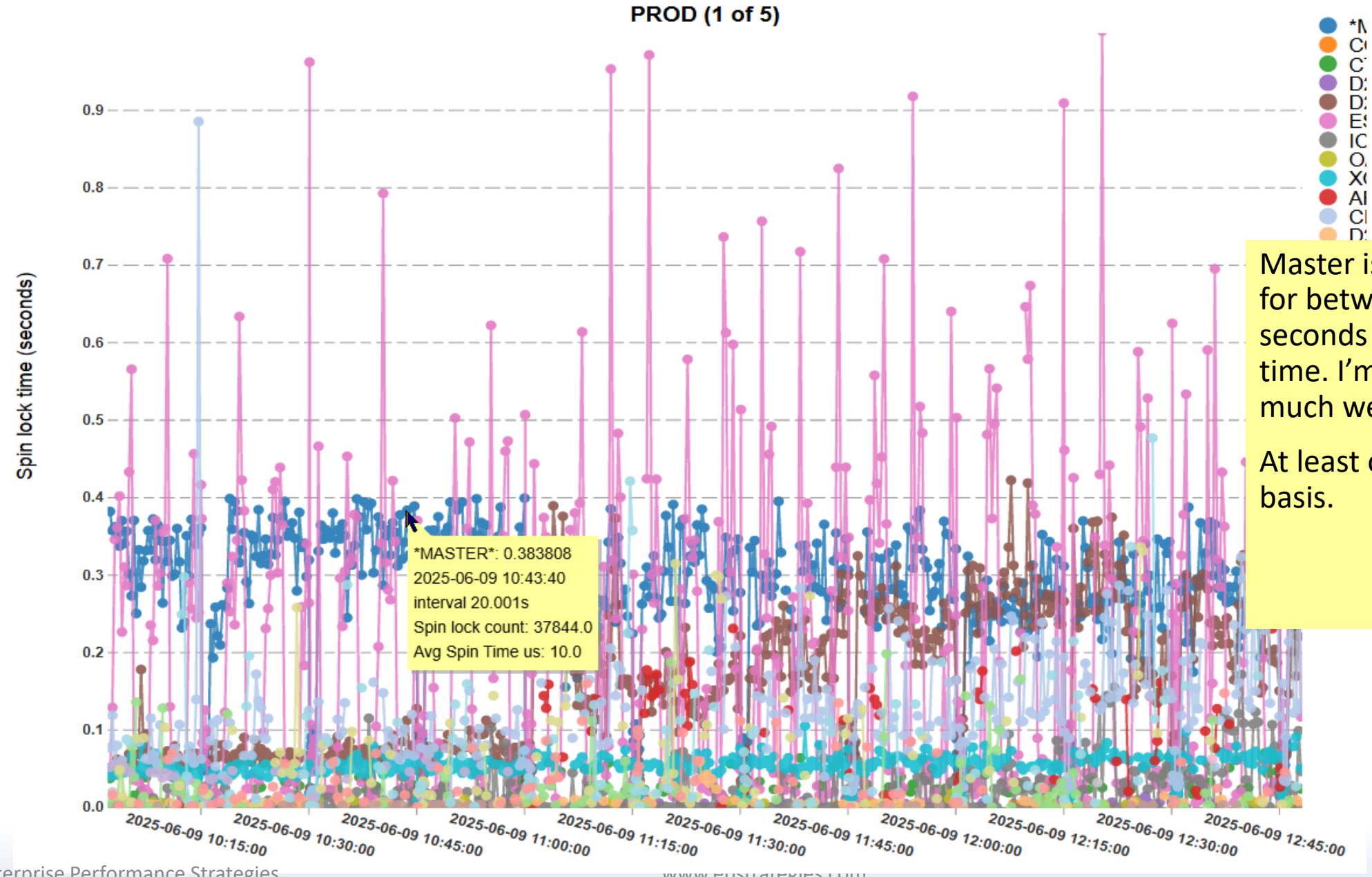
PROD (1 of 5)



Clearly there are some address spaces that regularly have higher CPIs than most others. This is probably more the nature of the work than a particular problem to worry about.

Address Space Spin Lock Time

When Recorded on HFTS Interval



Master is in spin locks for between 0.3 and 0.4 seconds around this time. I'm not sure how much we really care.

At least on a regular basis.

Summary

- SMF 98.1 does include interesting and potentially useful details about individual address spaces
- However, only some address spaces recorded per interval, so don't interpret a lack of data as a lack of activity!
- Record the 98s, ideally with a 5 second interval
  - Subtype 1 data generally under 500MB/system/day from what we've seen
- Even more data available if you enable WIC
  - Which may be an extra cost (if you don't license RMF)
  - May be an extra 2-3 GB of SMF data per system per day
  - We haven't really looked at this data yet