Understanding and Measuring Warning Track on z/OS

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!)

Warning Track Interrupts (WTIs) were first introduced a number of years ago on the zEC12. WTIs are used to maximize physical CPU utilization and improve response time by providing a communication method between the PR/SM hypervisor and z/OS. Essentially, PR/SM is warning z/OS that it is about to have a physical processor un-dispatched, and this gives z/OS an opportunity to remove a unit of work from the processor before the actual interrupt occurs. During this webinar, ***Scott Chapman*** will further explain Warning Track. Scott will also review the SMF measurements and whether those measurements are useful in understanding and managing your LPARs.

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

z/OS Performance workshops available

During these workshops you will be analyzing your own data!

- WLM Performance and Re-evaluating Goals
  - February 19-23, 2024
- Parallel Sysplex and z/OS Performance Tuning
  - August 20-21, 2024
- Essential z/OS Performance Tuning
  - September 16-20, 2024
- 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!

Like what you see?

- Free z/OS Performance Educational webinars!
  - The titles for our Fall 2023-2024 webinars are as follows:
    - ✓ *LPAR Configurations to Avoid*
    - ✓ *How Different are High, Medium, and Low Pool Processors?*
    - ✓ *CPU Critical: A Modern Revisit of a Classic WLM Option*
    - ✓ *Mainframe Efficiency at High Utilizations (Bob Rogers)*
    - ✓ *I/O, I/O It's Home to Memory We (Should) Go*
    - ✓ *30th Anniversary of WLM : A Retrospective and Lessons Learned*
    - ✓ *Mainframe Efficiency at High Utilizations (presented by Bob Rogers)*
    - *Understanding and Measuring Warning Track on z/OS*
    - *30th Anniversary of Parallel Sysplex - A Retrospective and Lessons Learned*
    - *Batch Initiators – WLM Managed or JES Managed?*
    - *AI on Z: Exploring Common AI Terms on System Z*
    - *Analyzing 'Per CPU' Utilizations*
    - *AI on Z: Exploring new SMF Measurements*
  - Let me know if you want to be on our mailing list for these webinars

- If you want a free cursory 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

Agenda

- What is Warning Track?
- What measurements do we have?
- What can we deduce from these measurements?

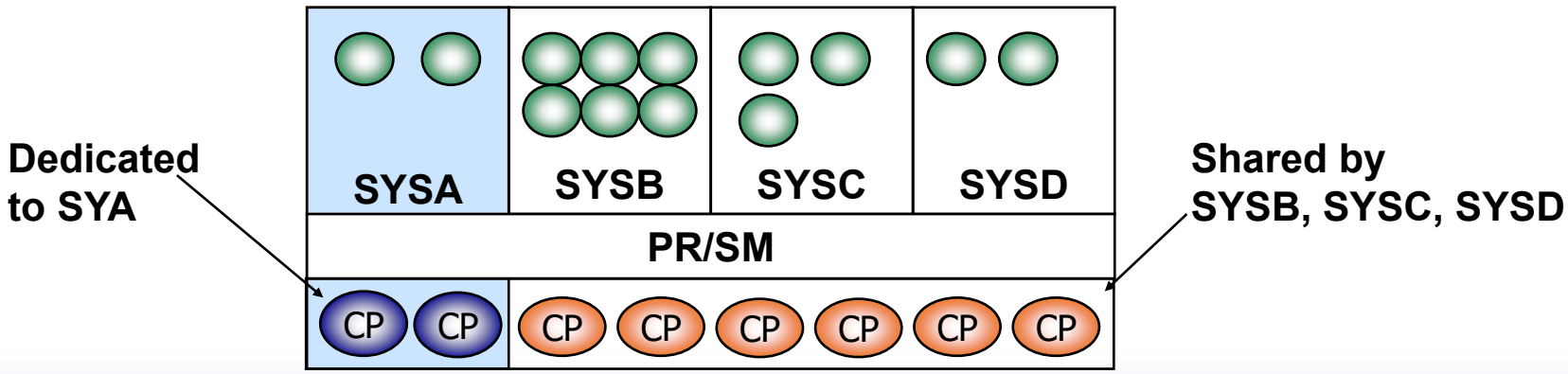
What is Warning Track

Background: Logical and Physical CPUs

- Processor = CP = CPU = GCPU or zIIP or any other processor type
  - All the same bit of silicon: a core on a physical chip
- You pay for a certain number of physical processors (CPs)
  - A processor can only be processing one stream of instructions at a time
    - Absent SMT, which don't apply to GCPs and which we're not going to discuss here
- You define LPARs, each with a certain number of logical, shared CPs
  - For each LPAR Logical CPs <= physical CPs, although can have reserved CPs
  - Most machines have multiple LPARs
- z/OS dispatches work to its (logical) CPs
- PR/SM dispatches logical CPs to physical CPs
  - A logical CP can't do any work when it's not dispatched to a physical CP

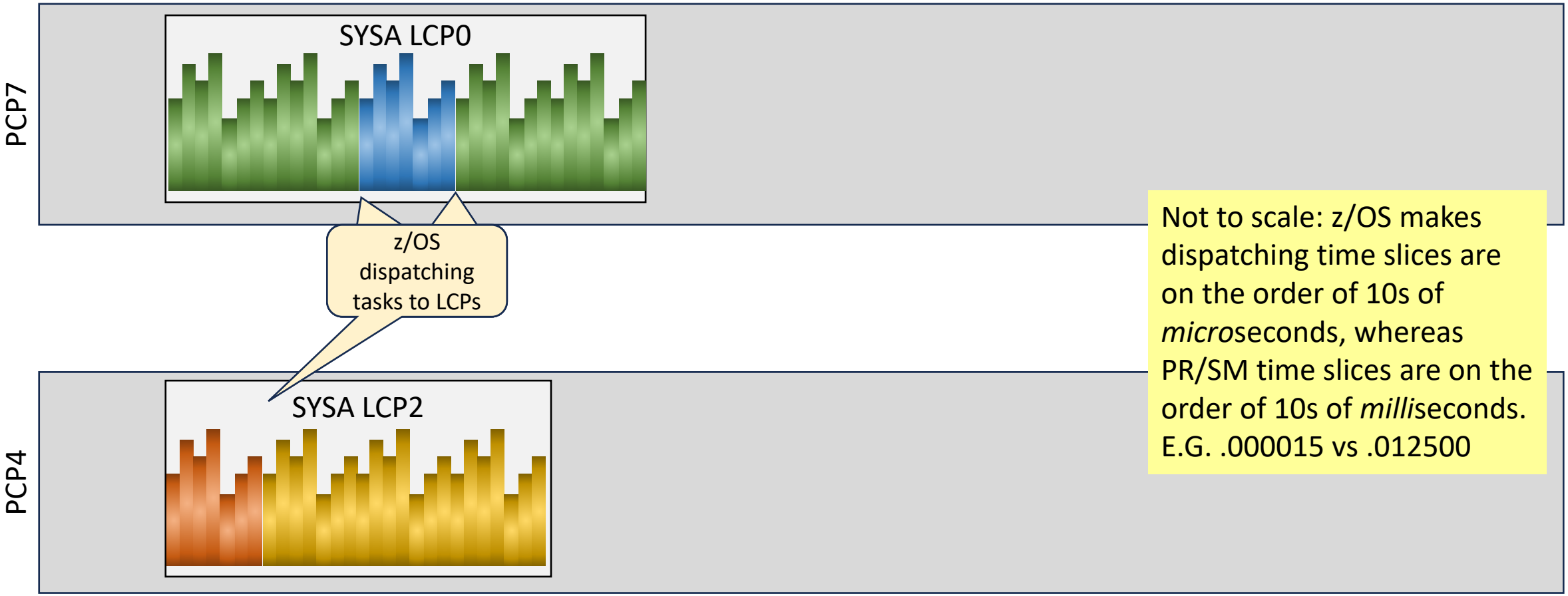
Dedicated CPs

- The prior slide is typical for z/OS LPARs: vast majority use shared CPs
- Can also use dedicated CPs (more common for ICF LPARs)
- Dedicated CPs represent CPs not shared between LPARs
- Dedicated CPs assigned directly to physical CPs
- Dedicated CPs reduces available shared CPs
  - So if you paid for 8 physical CPs, and have an LPAR with 2 dedicated CPs, that leaves 6 physical CPs for the remaining LPARs



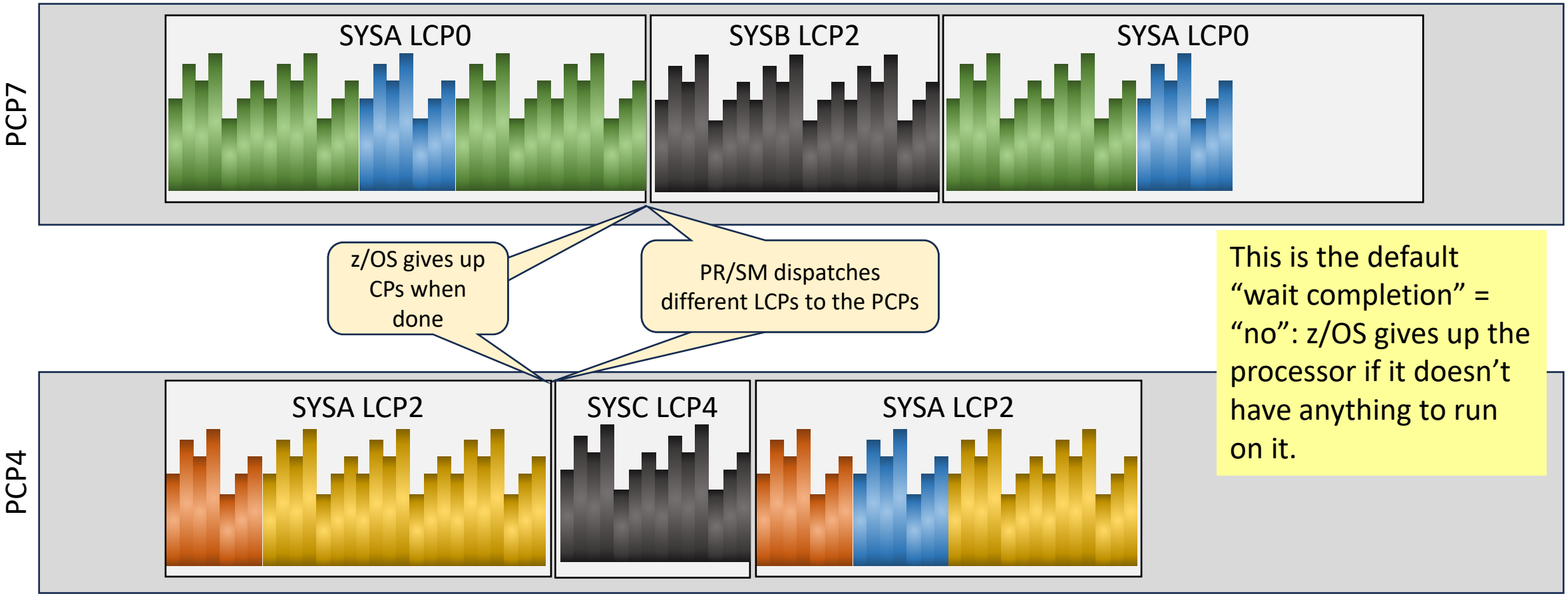
Unusual configuration for illustrative purposes

PR/SM Dispatching LCPs



Not to scale: z/OS makes dispatching time slices are on the order of 10s of *micro*seconds, whereas PR/SM time slices are on the order of 10s of *milli*seconds. E.G. .000015 vs .012500

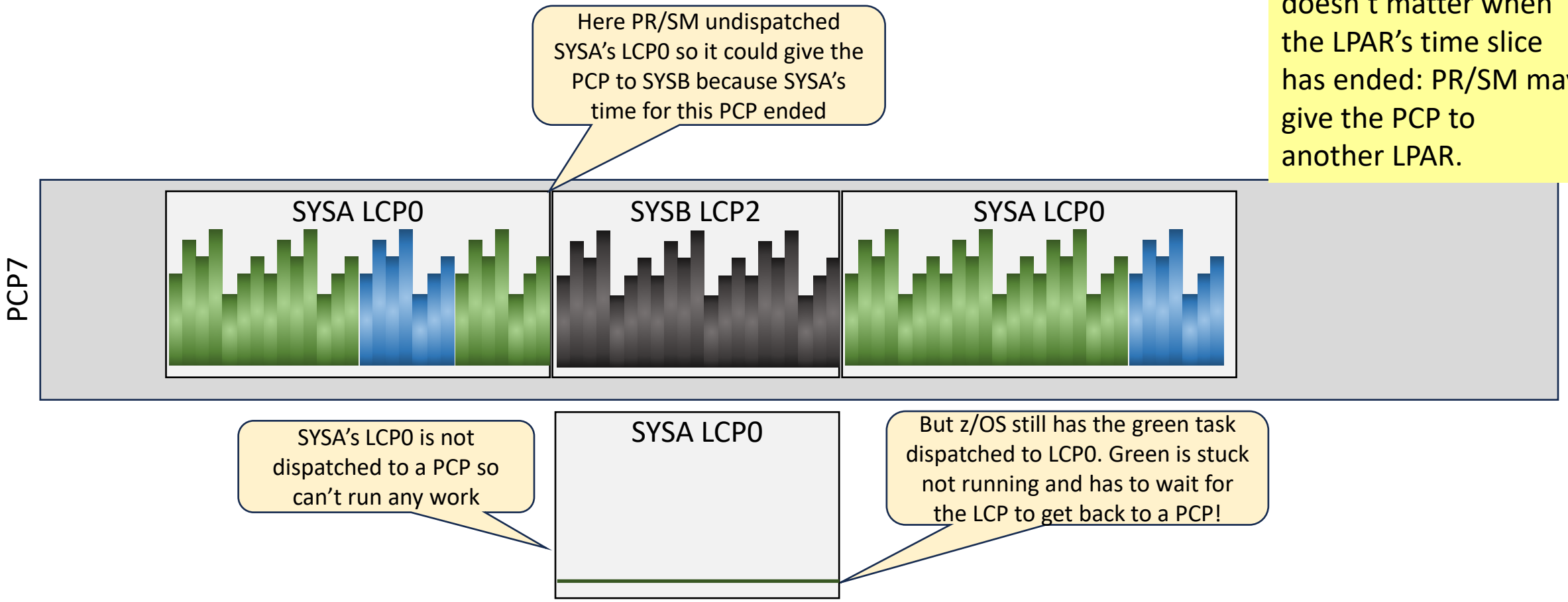
PR/SM Dispatching LCPs



This is the default "wait completion" = "no": z/OS gives up the processor if it doesn't have anything to run on it.

What if z/OS task wasn't done?

Wait completion doesn't matter when the LPAR's time slice has ended: PR/SM may give the PCP to another LPAR.



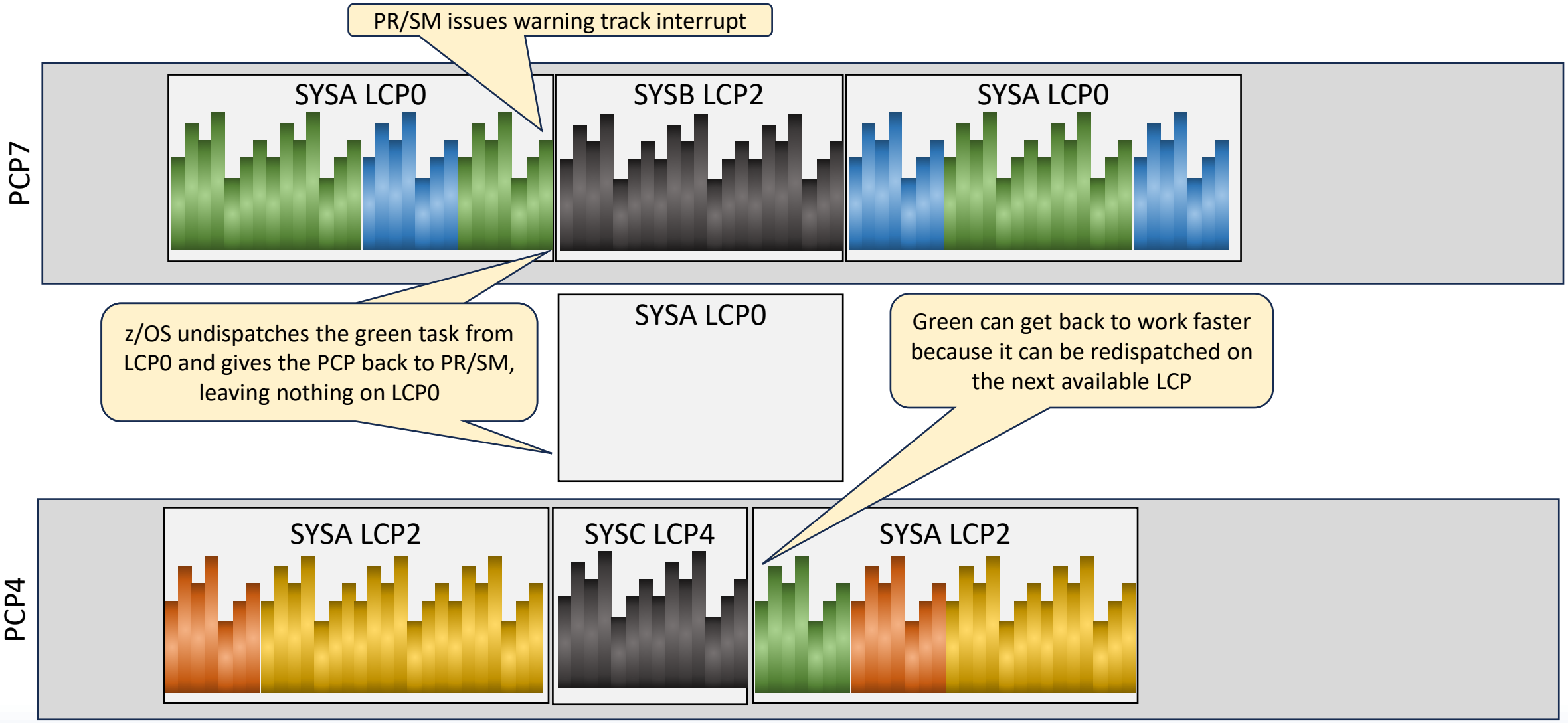
Involuntary Wait Issue

- When PR/SM steals an PCP from a z/OS LPAR when z/OS is still actively using it, the active task remains dispatched to the logical processor but is effectively suspended because it has no hardware to run on
- Note that with HiperDispatch this generally would only be expected to happen for Vertical Medium and Low processors
  - Vertical Highs are quasi-dedicated to the LPAR so if the LPAR's time slice ended but still has demand PR/SM would be expected to give the PCP back to the LPAR
- For Vertical Medium and Low processors, the PR/SM dispatch interval is between 12.5 and 25ms (often 12.5)
  - This can be a long time for a task to be involuntarily stranded
  - Worse: VLs might not come back for seconds (or longer) if they get parked
  - Can be especially painful if an important task gets stuck this way!

Warning Track Benefit

- Starting with the zEC12, PR/SM issues a warning track interrupt (WTI) to z/OS that it's about to take away the processor
- z/OS gets a grace period to un-dispatch the running task from the LCP and return the PCP to PR/SM
  - z/OS can then redispatch the task to an active LCP
  - "Successful" Warning Track Interrupt
- If z/OS doesn't return the processor in time, PR/SM takes it anyways
  - "Unsuccessful" Warning Track Interrupt
- Goal is to avoid having work hung on an LCP that's not going to get redispatched for some time

What if z/OS task wasn't done?



Warning Track Measurements

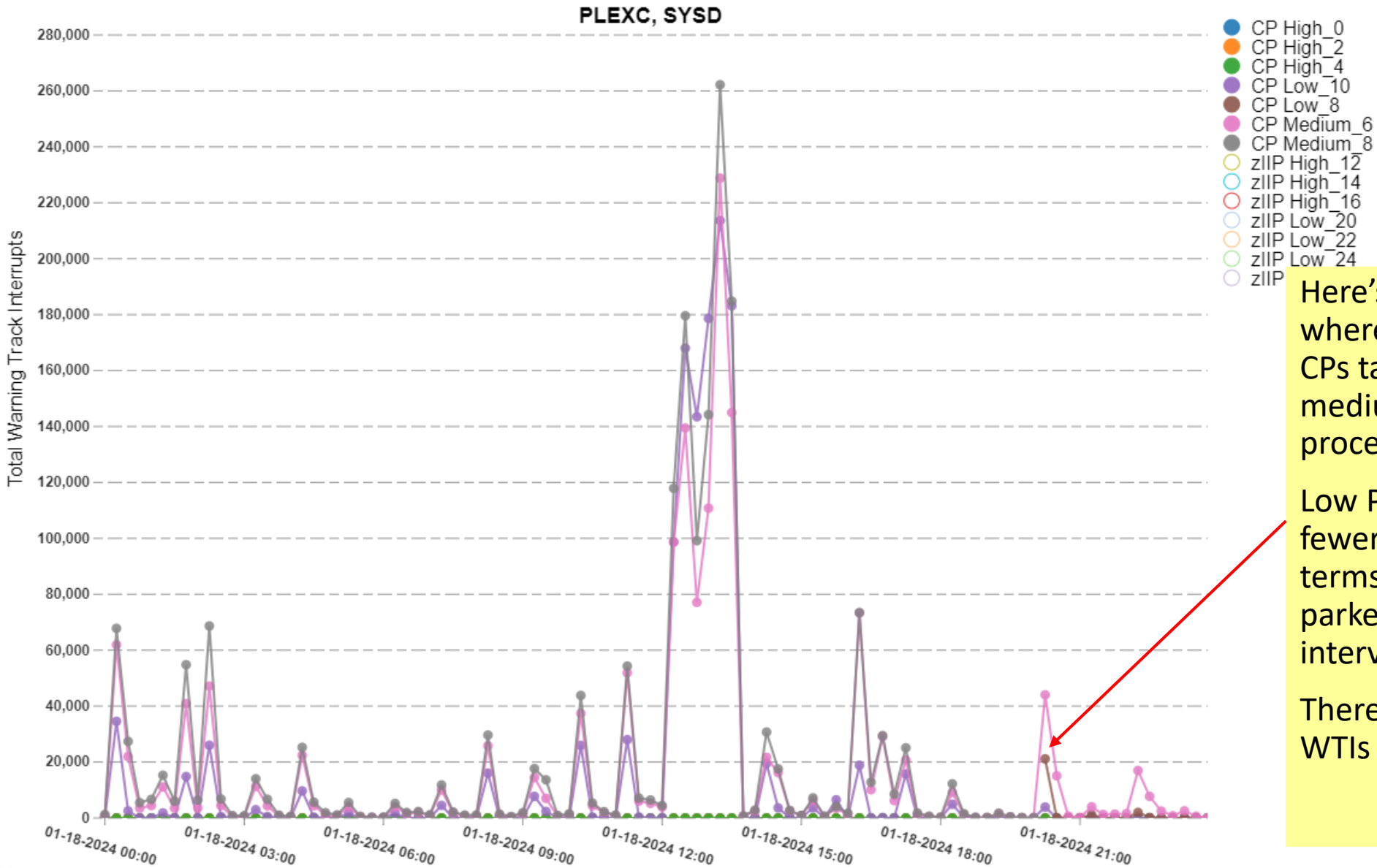
SMF Fields

- SMF 70 contains 3 fields of interest for warning track:
  - SMF70WTS – number of times PR/SM issued a WTI, and z/OS was able to return the processor within the grace period (“successful”)
  - SMF70WTU – number of times PR/SM issued a WTI, but z/OS was unable to return the processor within the grace period (“unsuccessful”)
  - SMF70WTI – time that a logical processor was yielded to PR/SM due to WTIs
- Those measurements are by logical processor

Expectations

- Expect no or very few WTIs for high-polarity processors
- Expect more WTIs per dispatched second for LCPs with lower weights
  - E.G. Expect more for vertical lows than vertical mediums
- Expect more WTIs when machine busier
- Expect most WTIs to be successful
- Expect total time yielded to be relatively low
  - Expect each successful yield to be small portion of the PR/SM time slice
  - Expect interrupts to be less than 80 / dispatch second

Warning Track Interrupts Per CPU

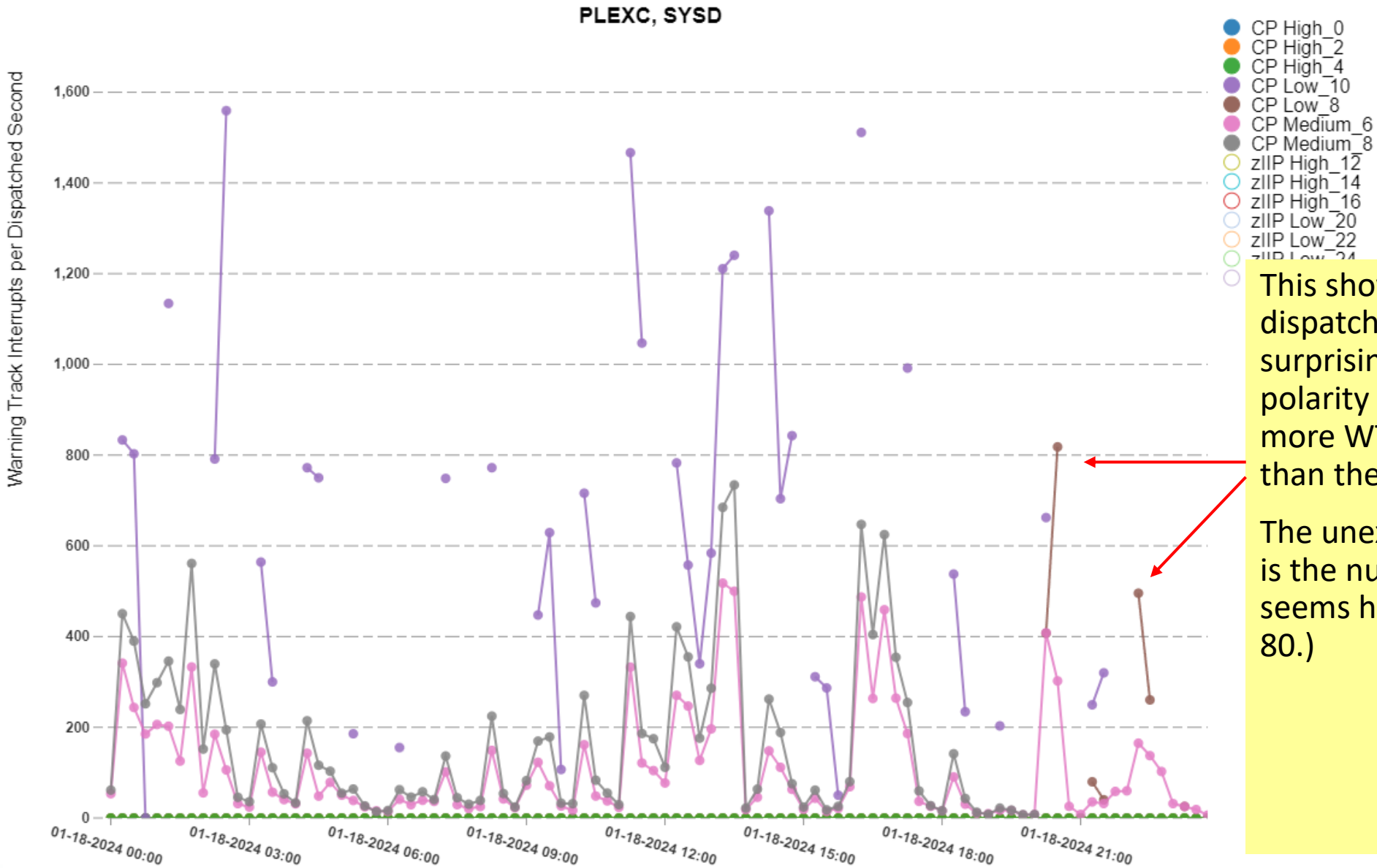


Here's a typical example where the High Polarity CPs take no WTIs, and the medium and low pool processors do.

Low Polarity CPs may have fewer WTIs in absolute terms because they may be parked for much of the interval.

There clearly were more WTIs right after noon.

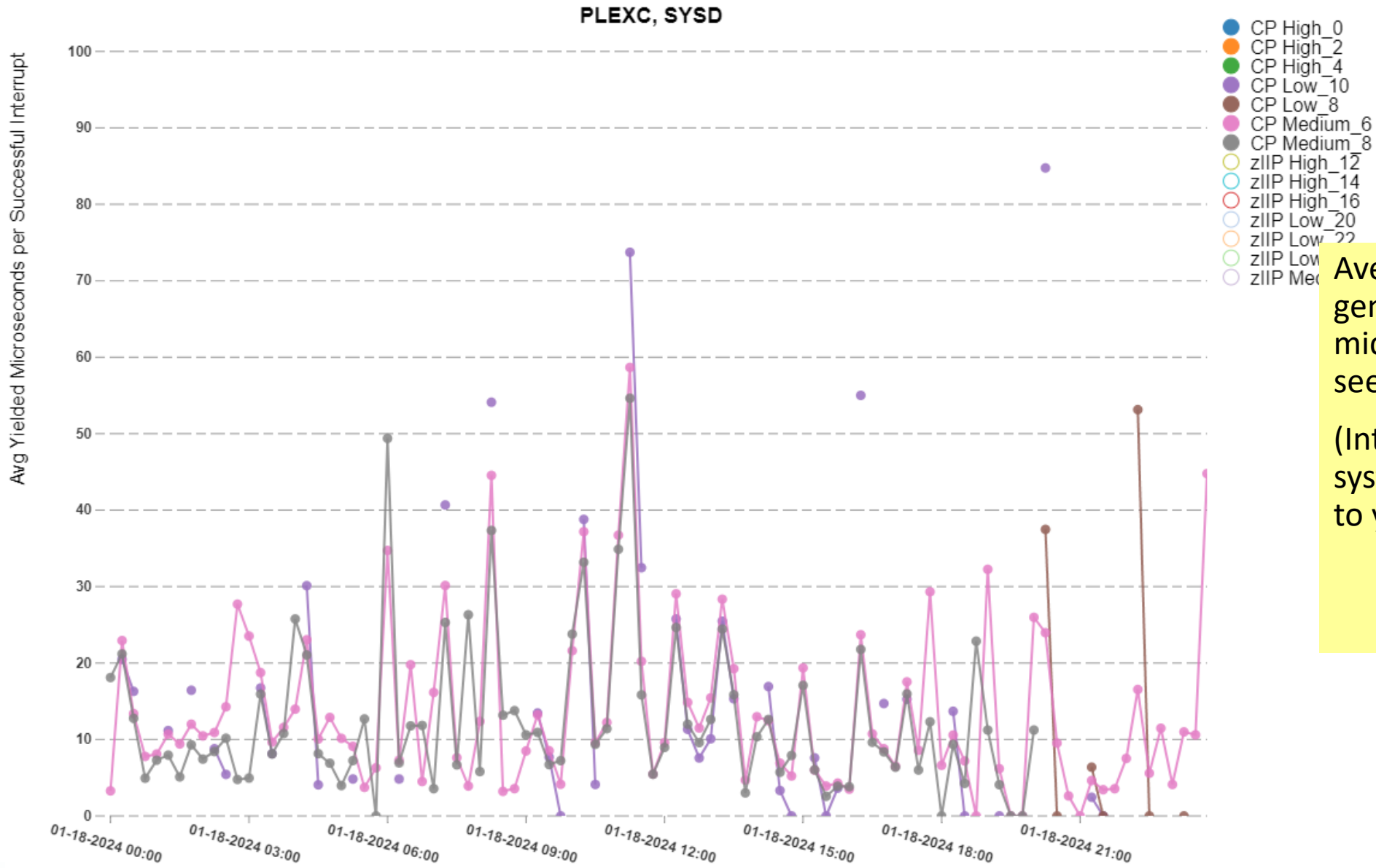
Warning Track Interrupts per Dispatched Second



This shows the WTIs per dispatched second and less surprisingly, the low polarity processor has more WTIs per second than the medium.

The unexpected part here is the number per second seems high. (Expected <= 80.)

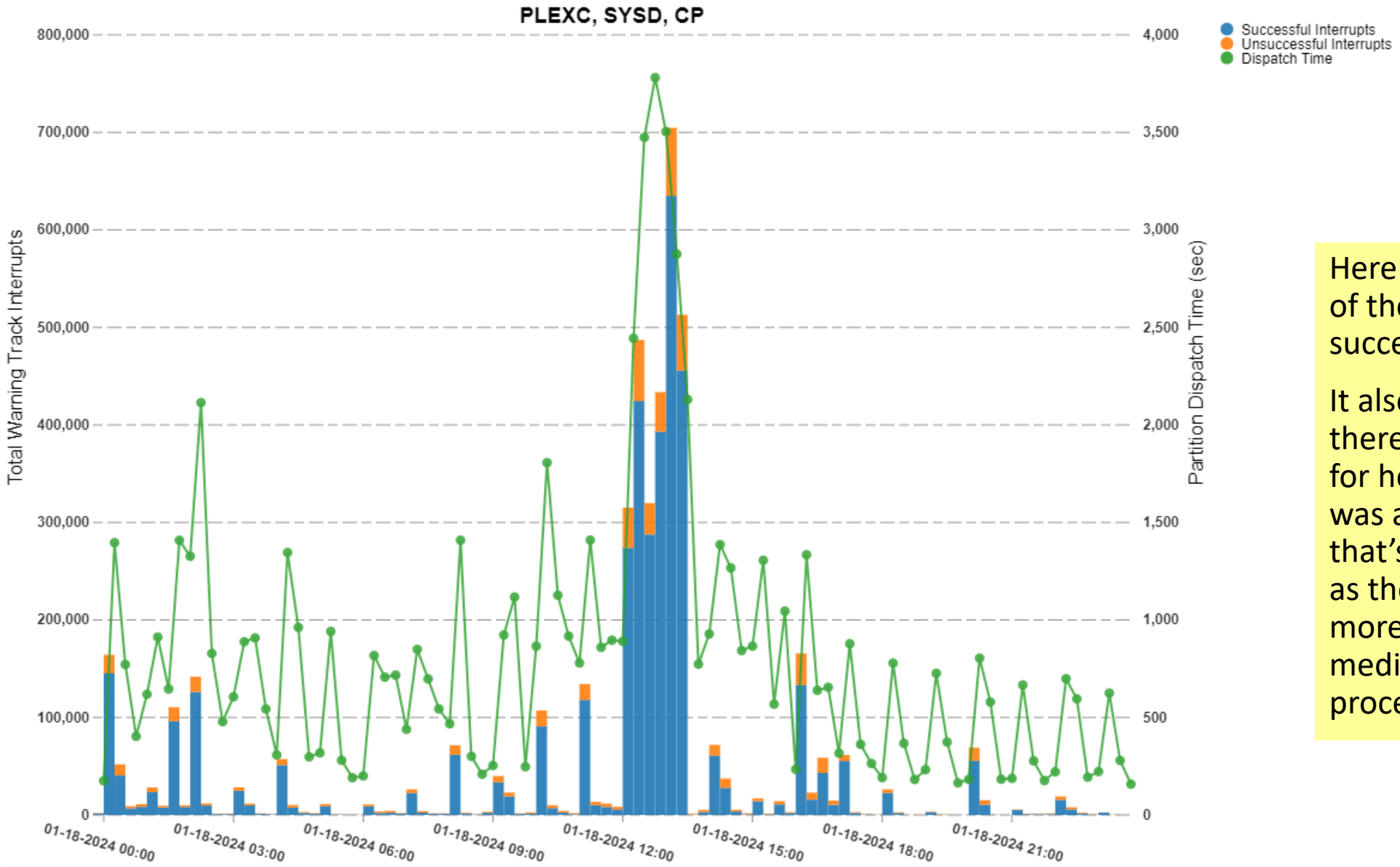
Warning Track Interrupts Avg Yielded Microseconds



Average yield time was generally under 40 microseconds, which seems reasonable.

(Interestingly, for this system, the zIIPs tended to yield more time.)

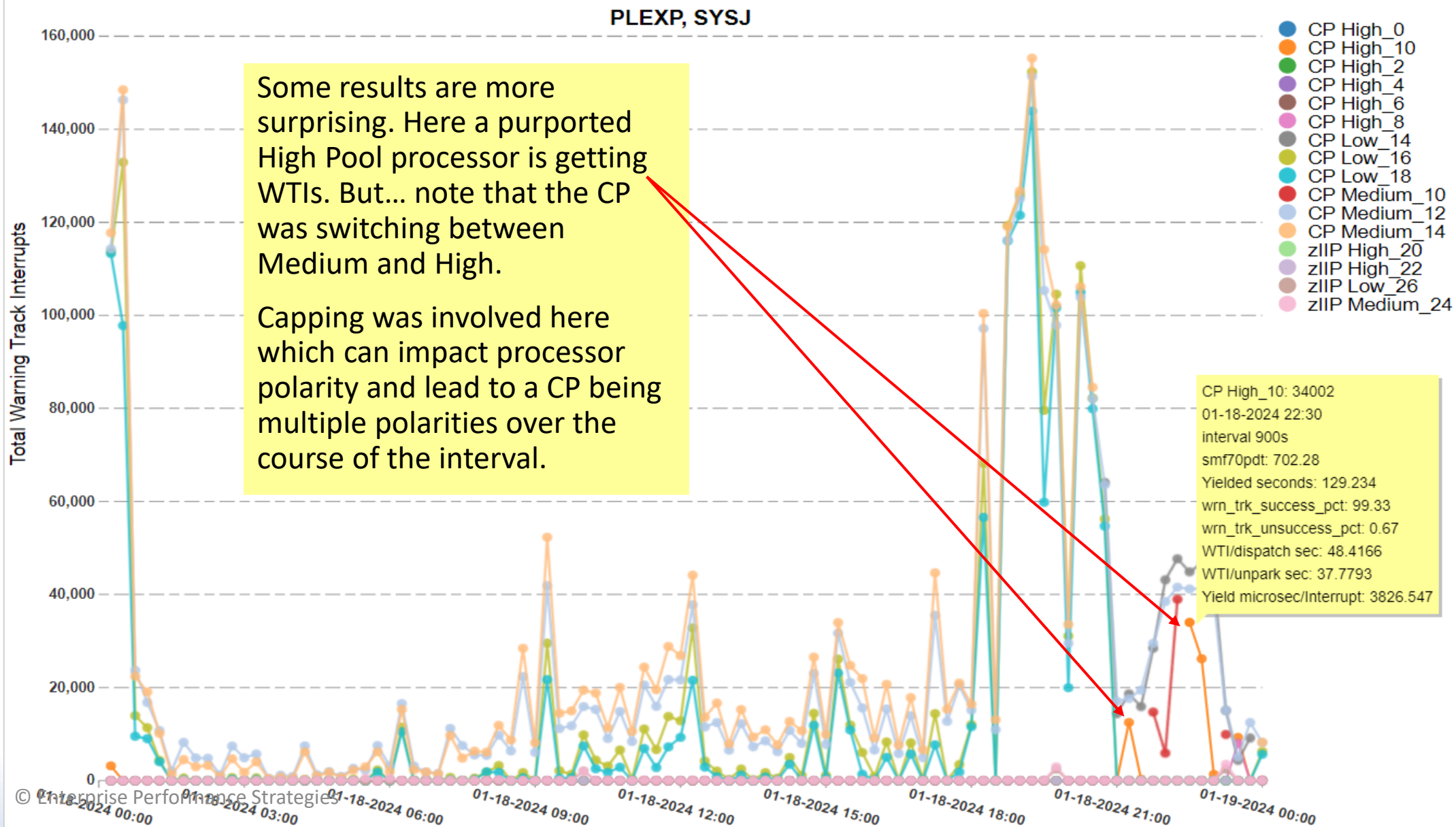
Warning Track Interrupts vs. Dispatch Time



Here we see that most of the WTIs were successful.

It also appears that there was a correlation for how busy the LPAR was and the WTIs. But that's probably because as the LPAR got busier more work was run the medium and low pool processors.

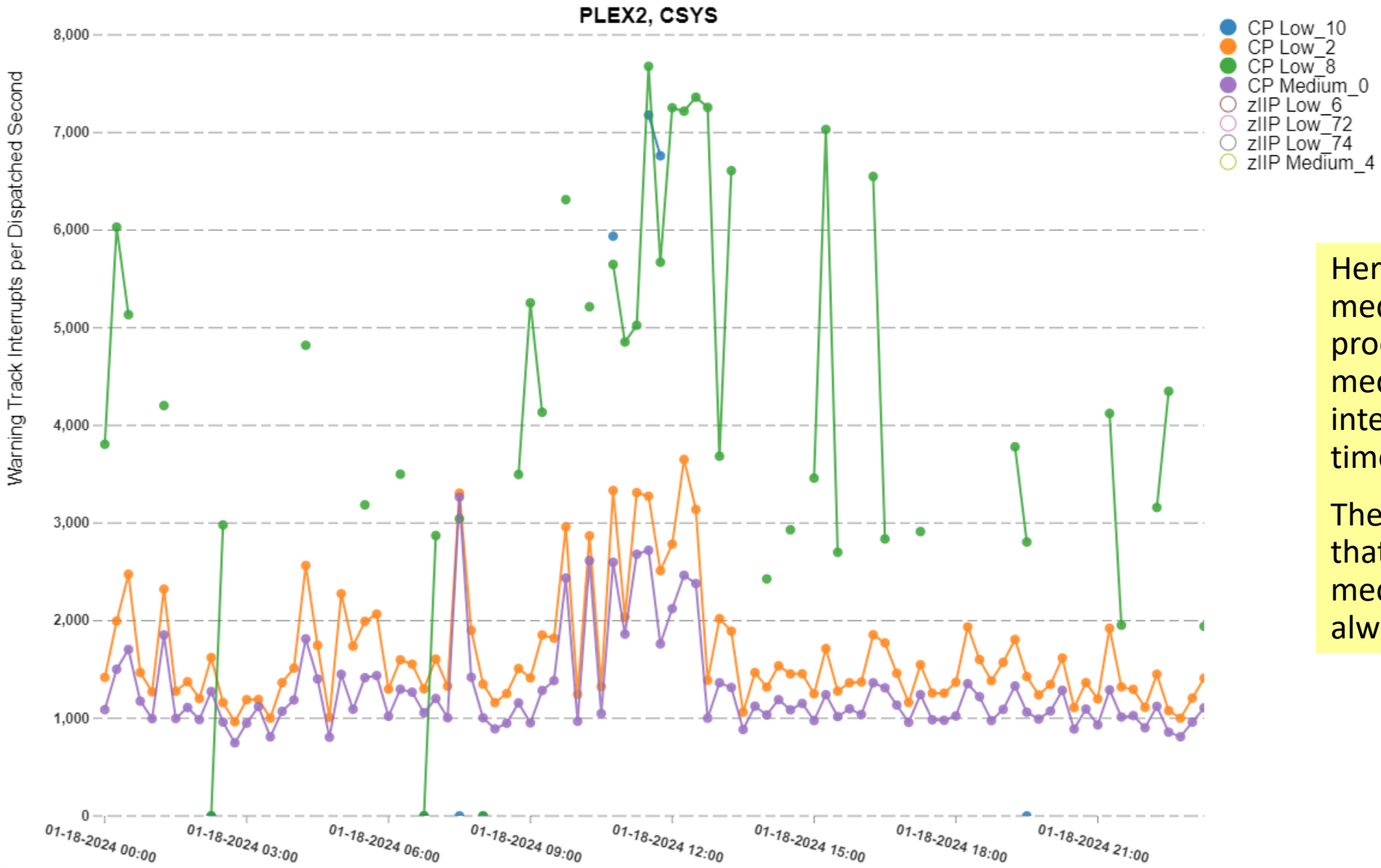
Warning Track Interrupts Per CPU



Some results are more surprising. Here a purported High Pool processor is getting WTIs. But… note that the CP was switching between Medium and High.

Capping was involved here which can impact processor polarity and lead to a CP being multiple polarities over the course of the interval.

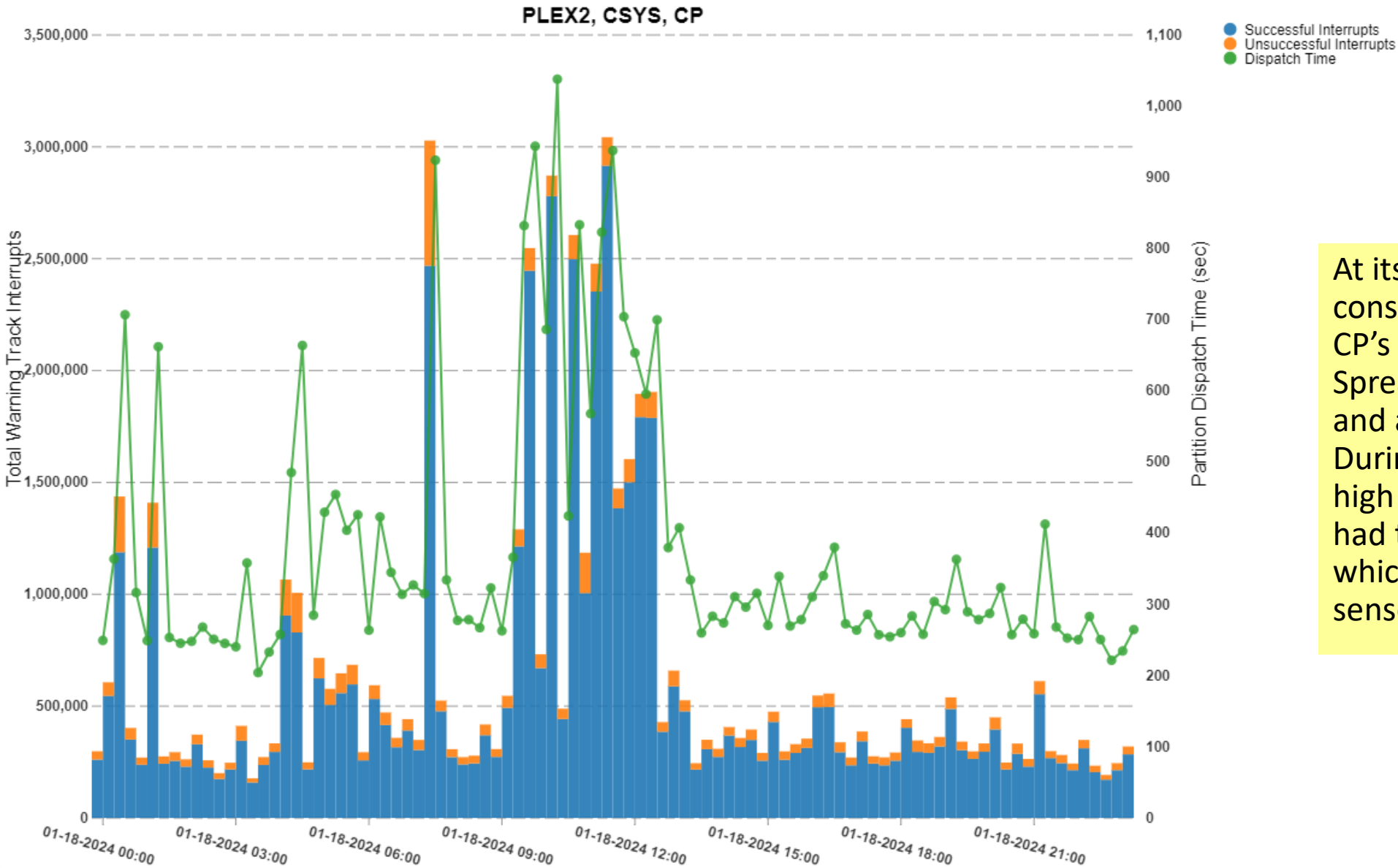
Warning Track Interrupts per Dispatched Second



Here this LPAR only has medium and low pool processors, and the medium is getting interrupted over 1000 times / second.

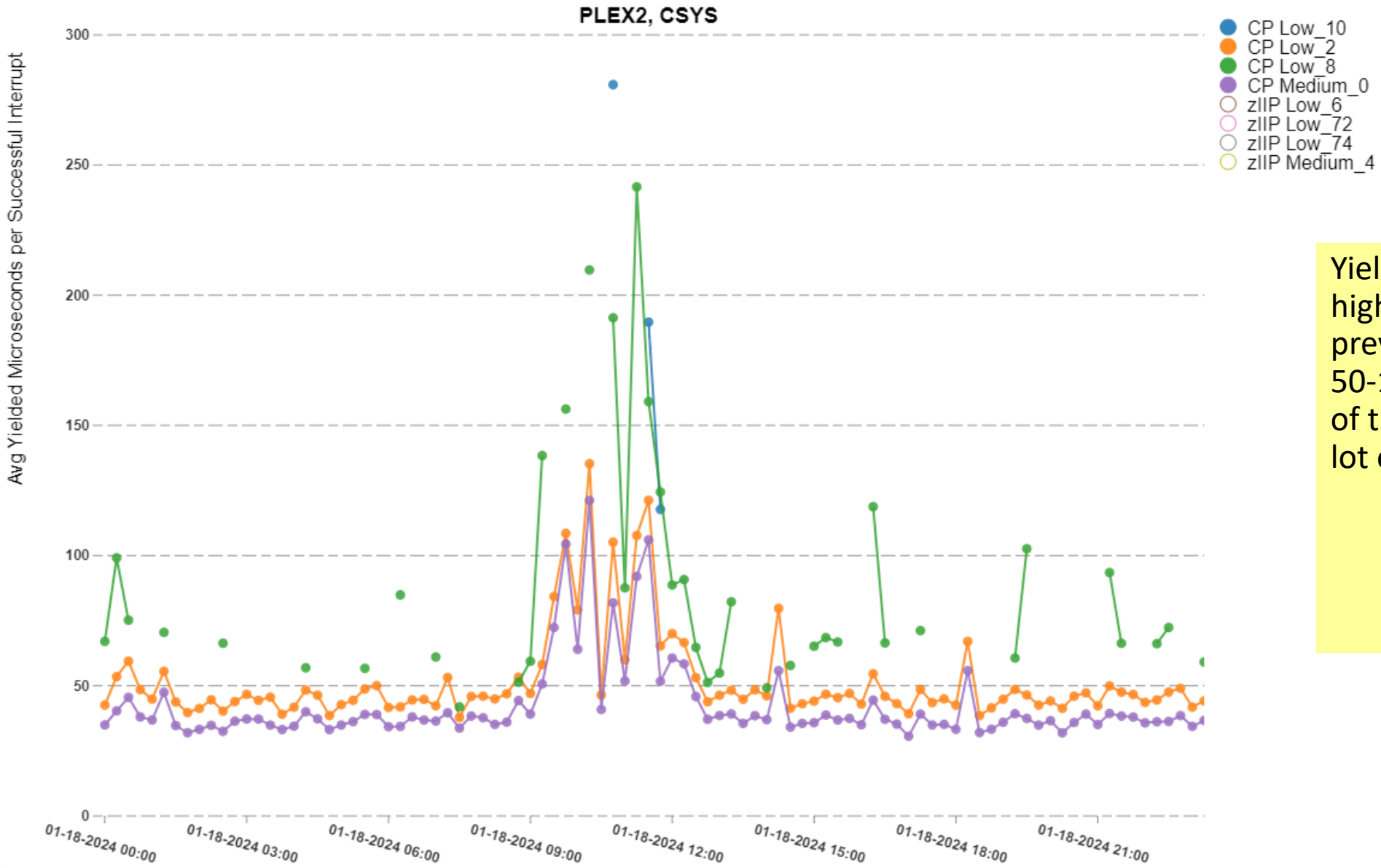
The low pool processor that is similar to the medium is probably the always unparked low.

Warning Track Interrupts vs. Dispatch Time



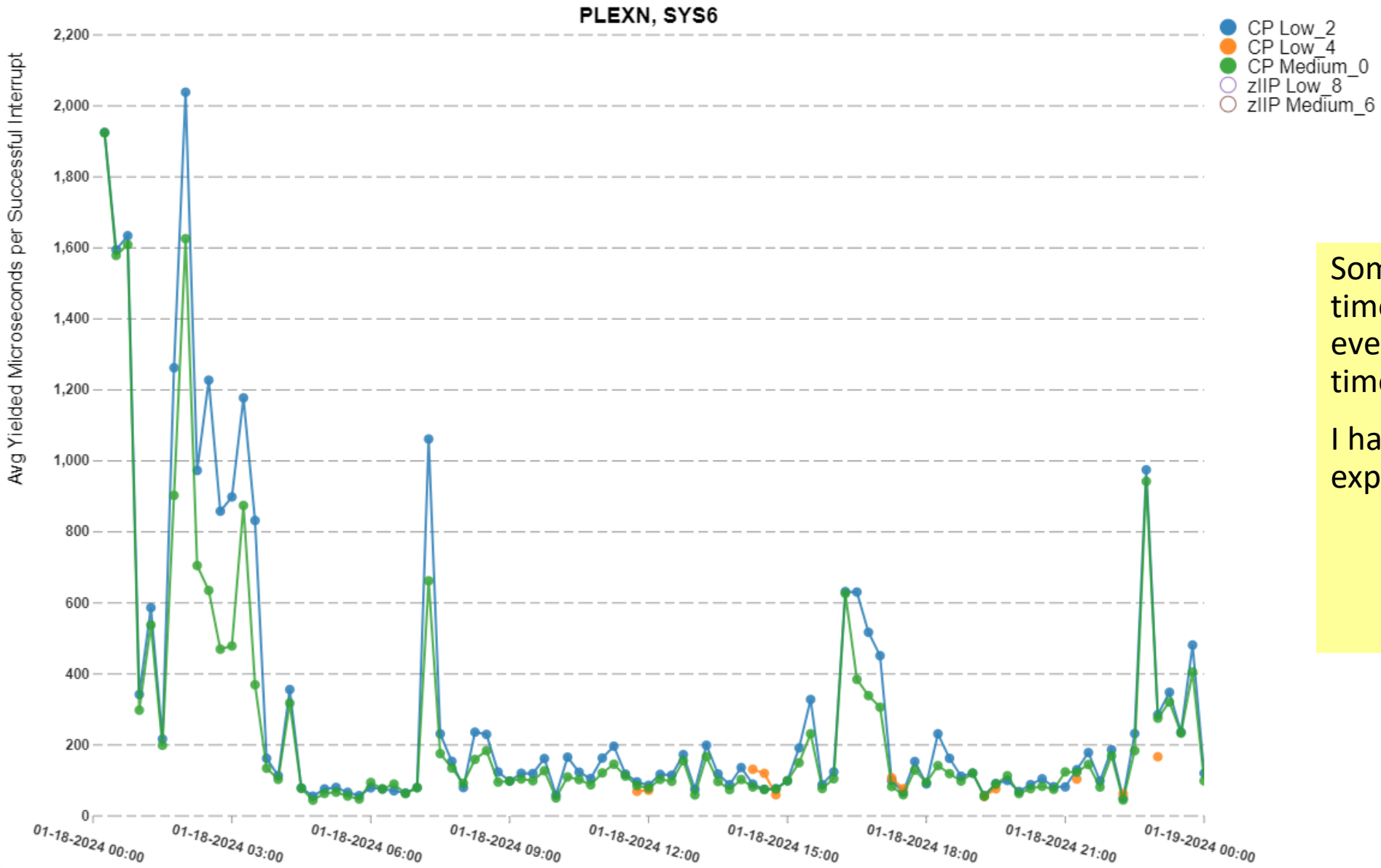
At its peak, the LPAR consumed over a full CP's worth of capacity. Spread across a medium and at least 2 lows. During these periods of high usage was when it had the most WTIs, which makes some sense.

Warning Track Interrupts Avg Yielded Microseconds



Yielded time seems higher than in the previous example. While 50-100µsec is not a lot of time to me, but it's a lot of time to a CPU.

Warning Track Interrupts Avg Yielded Microseconds



Sometimes the yielded time average can be even crazier. Here, at times, it's over 1ms!

I have no good explanation for this!

Burning questions

- How can there possibly be that many WTIs happening on medium and (especially) low polarity processors??
  - Maybe the numbers are just wrong?
  - Maybe PR/SM is doing something with the time slicing for mediums and lows that I haven't seen documented anywhere?
- Why do the yielded time averages sometimes seem excessive?
  - I'm more willing entertain the idea that this number might be inaccurate
  - Or could be related to the average unit of work size running in that system
- Do we care?
  - Everything is "working" so maybe not?
  - But it *seems* like this should be indicative of an inefficiency
    - But how much of an inefficiency?
    - And could we really do anything about it?

Recommendations

- Watch this space: I have an open to-do to find an explanation from IBM re. the WTI rate
  - I suspect (hope) this is a documentation issue more than a real issue
  - But if this turns out to be something important, we'll be sure to update you in a future webinar
- Adjusting weights to better match LPARs' requirements remains a best practice
  - Primarily to avoid having to run too much on low pool processors
  - In addition to potentially having worse cache hits, they may take more WTIs (leading also to worse cache hits)
  - Consider using automation to adjust weights at different times to match your processing requirements

Questions?