zIIP Capacity And Performance

Martin Packer, IBM

martin_packer@uk.ibm.com
Twitter/Facebook: martinpacker

Abstract

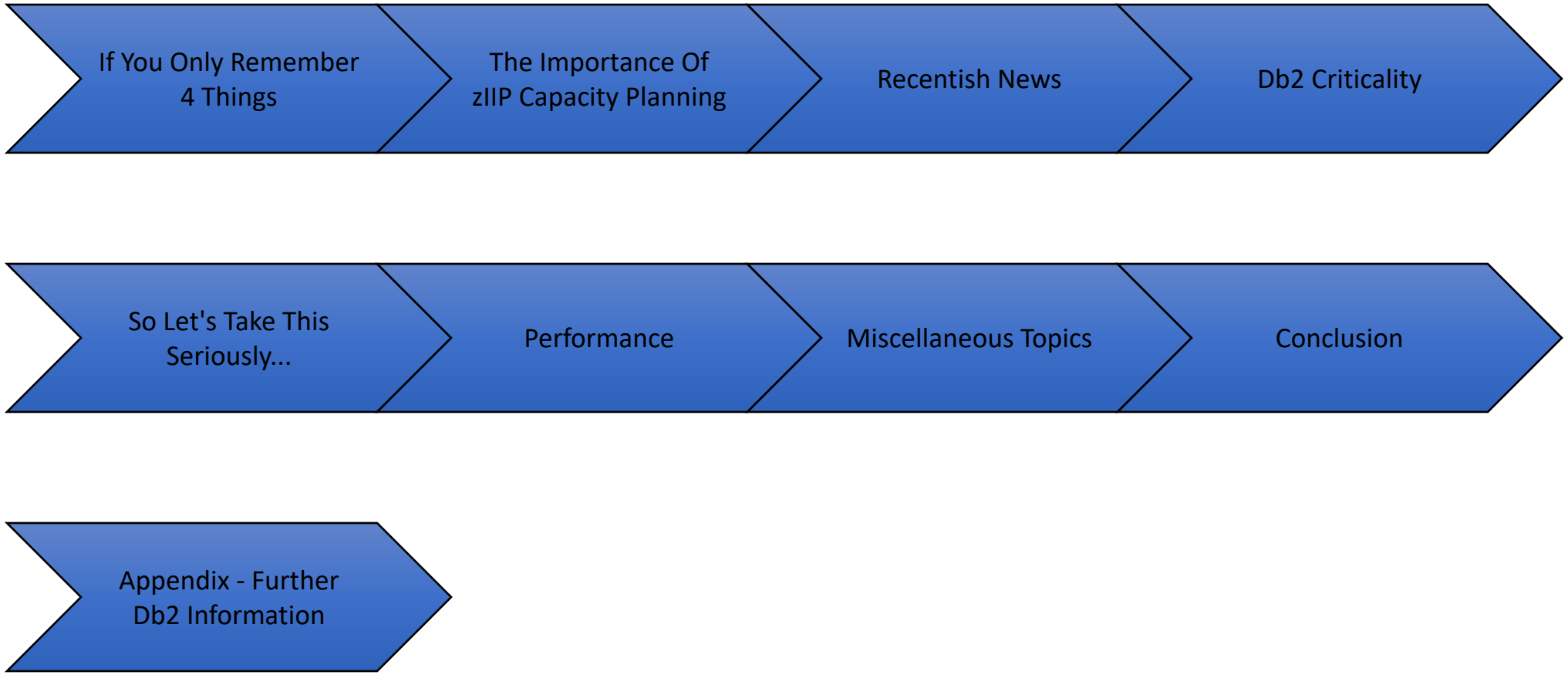
zIIP Capacity Planning tends to be neglected - in favour of General-Purpose Engines (GCPs). With recent enhancements to Db2 allowing you to offload critical CPU to zIIPs, and to get the most out of zAAP-on-zIIP, it's time to take zIIP Capacity Planning seriously.

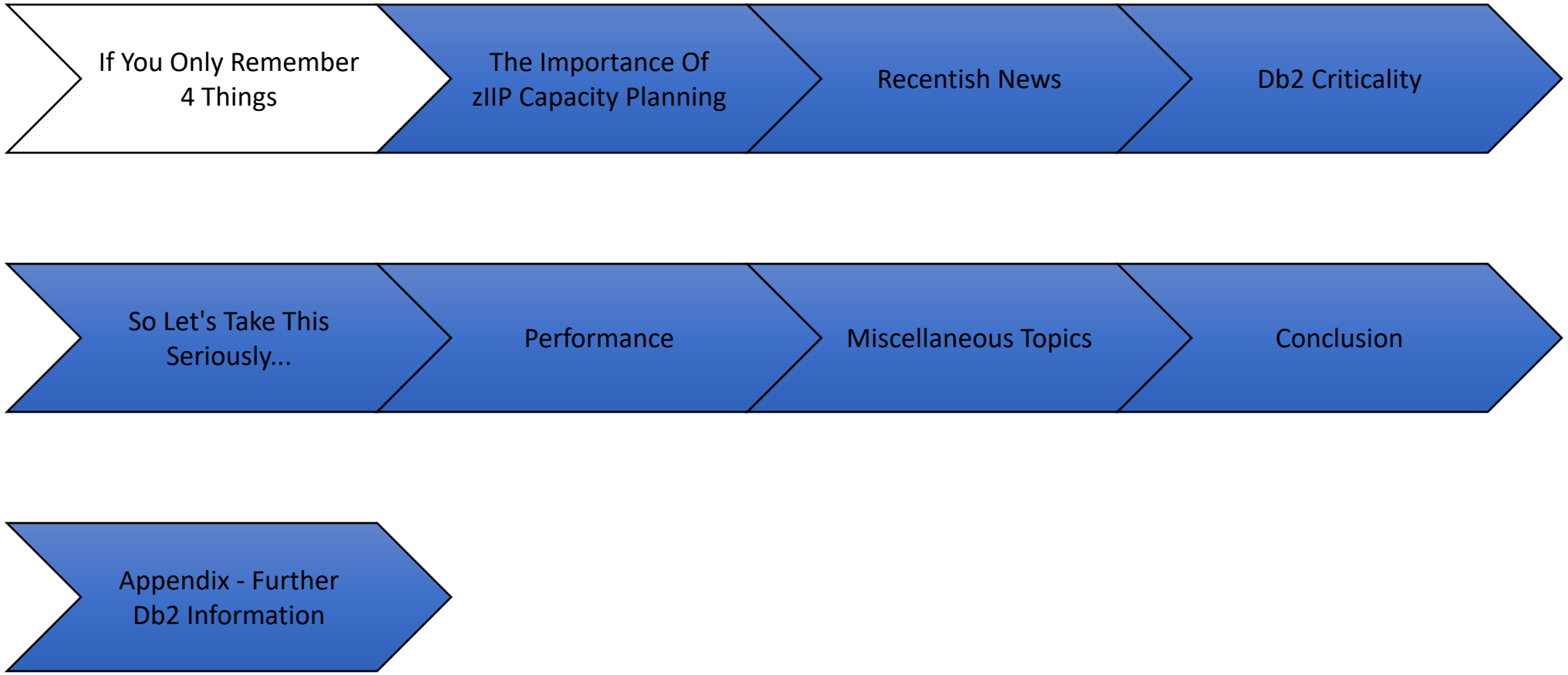
This presentation describes how to do zIIP Capacity Planning properly - with instrumentation and guidelines. It also covers some important Performance topics.

Note: Updated in 2018 for z13 SMT and numerous exploiting software updates

Note: Updated in 2020 for Performance considerations, System Recovery Boost, and zCX

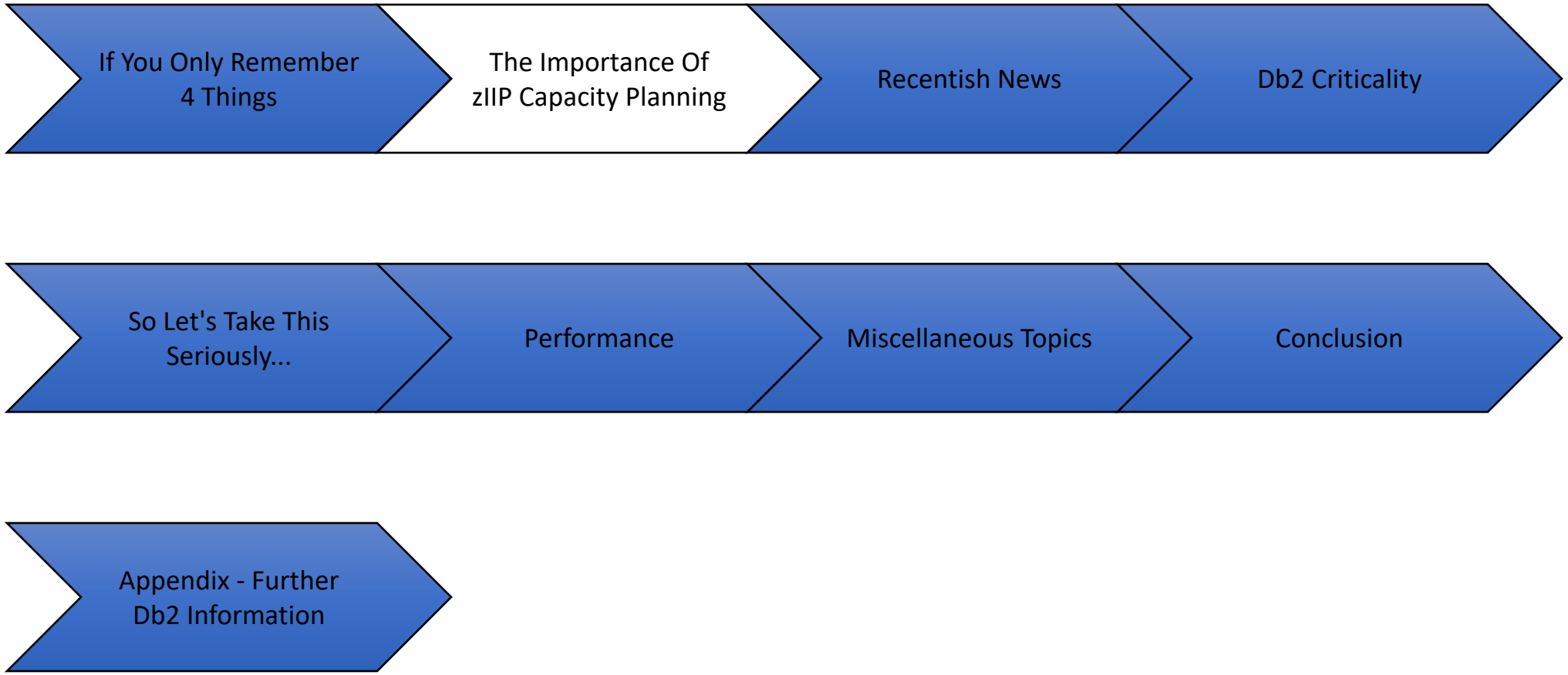
Topics


If You Only Remember 4 Things
The Importance Of zIIP Capacity Planning
Recentish News
Db2 Criticality
So Let's Take This Seriously...
Performance
Miscellaneous Topics
Conclusion
Appendix - Further Db2 Information


If You Only Remember 4 Things
The Importance Of zIIP Capacity Planning
Recentish News
Db2 Criticality
So Let's Take This Seriously...
Performance
Miscellaneous Topics
Conclusion
Appendix - Further Db2 Information

If You Only Remember 4 Things

- Capacity Planning for zIIPs has become much more important
- Pay especial attention to what % Busy represents “full”
- Measure zIIP potential and usage down to the address space level
- Consider LPAR configuration carefully for zIIPs

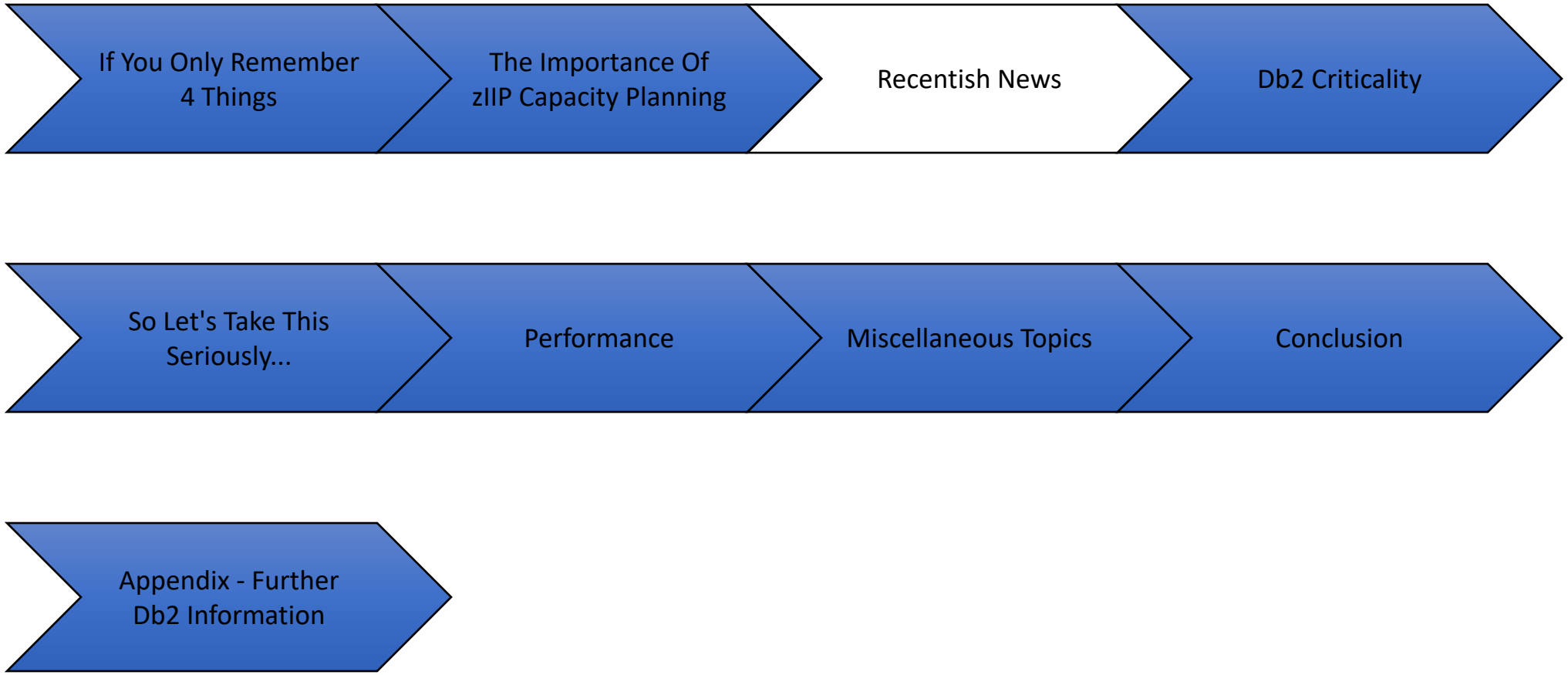

If You Only Remember 4 Things
The Importance Of zIIP Capacity Planning
Recentish News
Db2 Criticality
So Let's Take This Seriously...
Performance
Miscellaneous Topics
Conclusion
Appendix - Further Db2 Information

Why People Don't Do zIIP Capacity Planning

- zIIPs are much cheaper than general purpose processors (GCPs)
    - Though they run at full speed
    - Maintenance cost lower also
- zIIPs don't attract a software bill
- zIIP Utilization hasn't historically been all that high
- zIIP-eligible work can spill to GCPs at a pinch
- zIIPs are sometimes not the most carefully considered part of a purchase
- Ignoring zIIPs, ICFs, IFLs simplifies things to a single processor pool

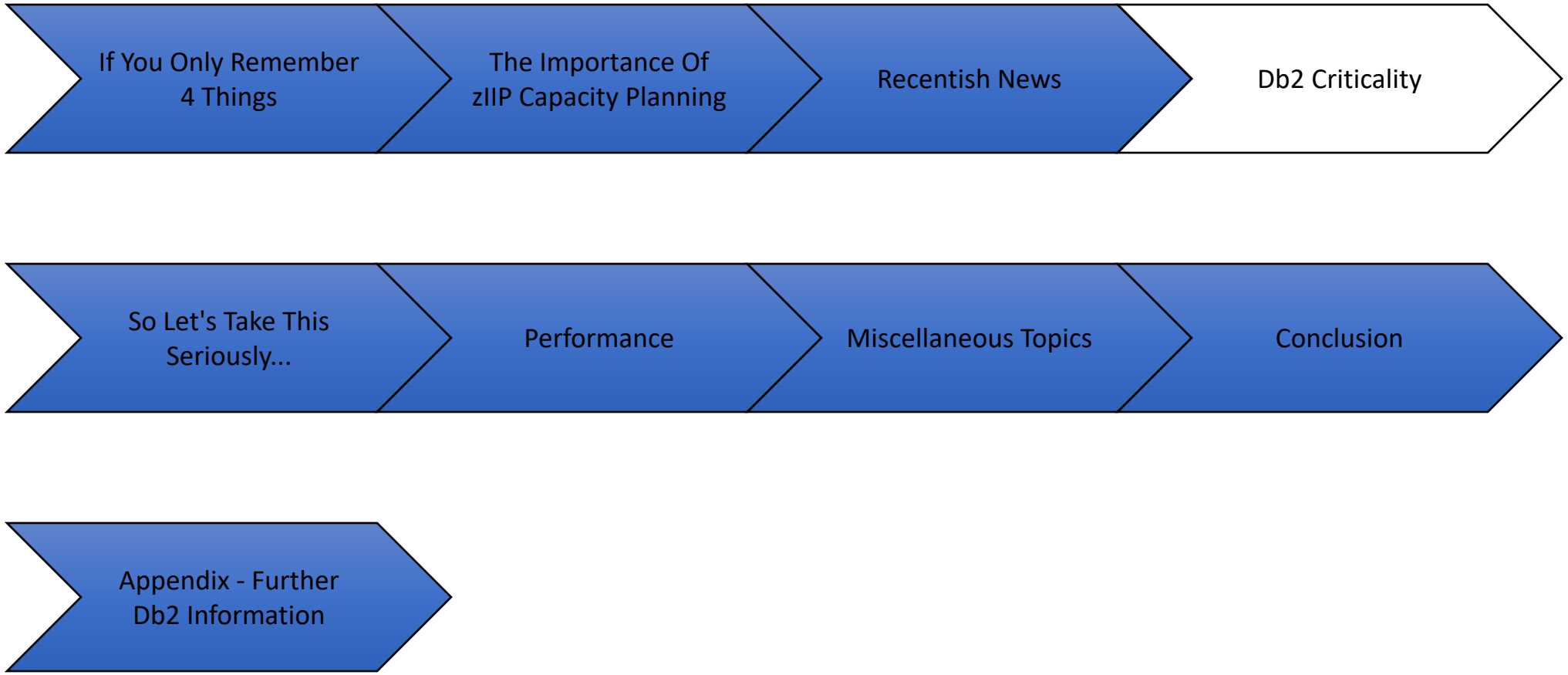
But Consider Why Many Customers Are Now Taking zIIP Seriously

- Spilling to a GCP costs money
- Not spilling to a GCP can cost performance & scalability
- zIIPs aren't free
- Shoving another processor drawer in can be difficult or impossible
    - And at end of processor life upgrades become impossible
- More software that exploits zIIP installed
    - Some more performance-critical than previous exploiters
- New and modernised applications increasingly use zIIP
- LPAR configuration for zIIPs affects their performance
- Failure to plan and provision capacity can lead to outages


If You Only Remember 4 Things
The Importance Of zIIP Capacity Planning
Recentish News
Db2 Criticality
So Let's Take This Seriously...
Performance
Miscellaneous Topics
Conclusion
Appendix - Further Db2 Information

Recent (& Now Not-So-Recent) News

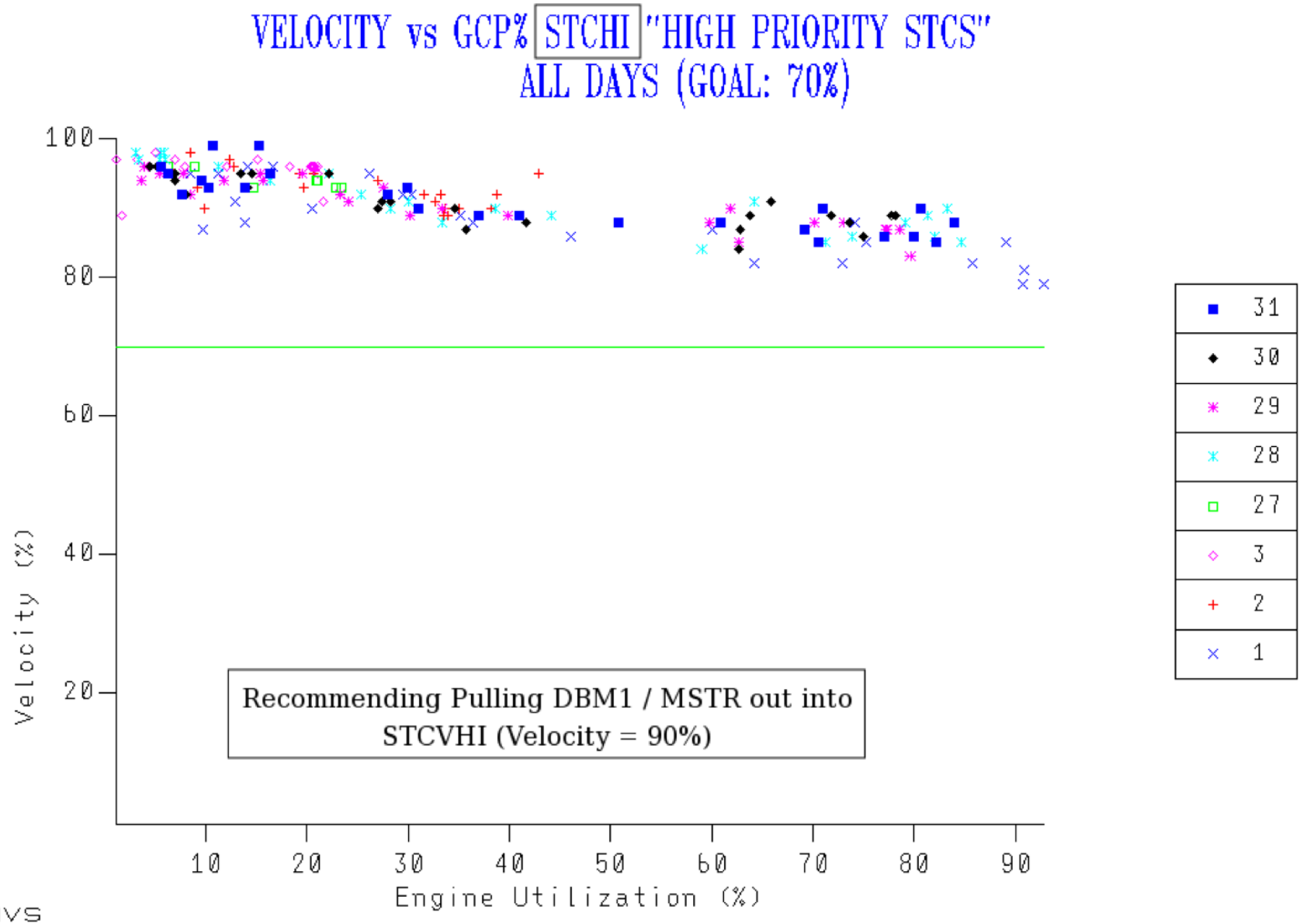
- zAAP On zIIP
  - Allows zAAP-eligible work to become zIIP-eligible
- 2:1 zIIP:GCP configuration rule
- Db2 Version 10 Deferred Write and Prefetch Engines
- Db2 Version 11 further exploiters
  - Log write and log read
- Db2 Version 12 yet more exploiters
- z13 zIIP SMT
- z/OS HonorPriority at Service Class Level (OA50845 / OA50953)
- z/OSMF Autostart in z/OS 2.3
  - Java but Liberty Profile
- z/OS 2.4 zCX Docker Container Extensions are zIIP-eligible
- Db2 High Performance Unload can sometimes use zIIPs
- z15 System Recovery Boost (SRB)
- Spark exploits zIIP


If You Only Remember 4 Things
The Importance Of zIIP Capacity Planning
Recentish News
Db2 Criticality
So Let's Take This Seriously...
Performance
Miscellaneous Topics
Conclusion
Appendix - Further Db2 Information

Db2 Version 10 Deferred Write And Prefetch Engines

- Substantial portion of DBM1 address space CPU
  - Changed by APAR PM30468 from MSTR address space
- 100% eligible for zIIP
- Very stringent performance requirements
  - Must not be delayed
  - Latency to cross over to GCP generally too high
- Substantially changes zIIP Capacity Planning rules
- **Treat DBM1 as a key address space and protect its access to CPU, including zIIP**
  - Set its Service Class to use CPU Critical
  - Likewise MSTR - especially with Version 11

Ensure Your Goals Protect Important Work & Understand The Role of zIIP In Goal Attainment

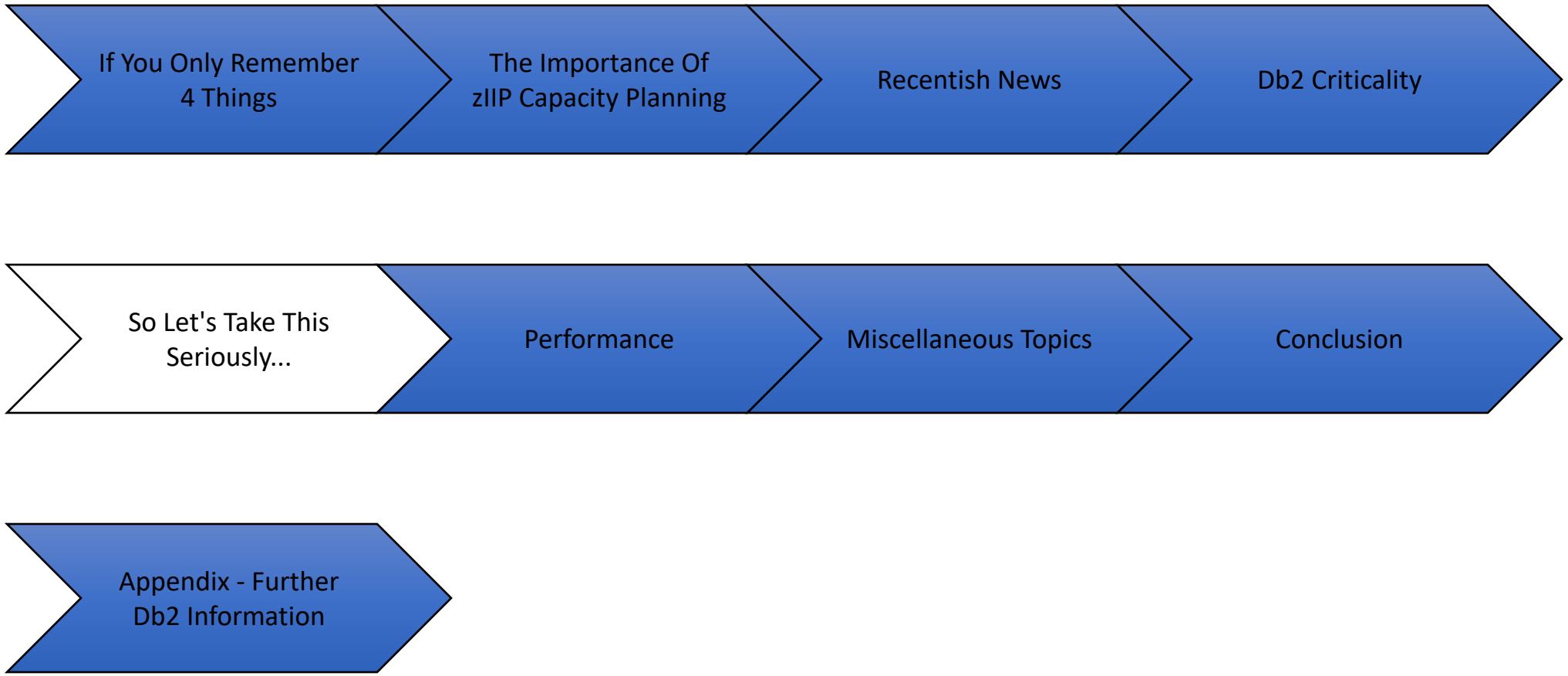


Db2 Version 11 Exploits zIIP Still Further

- Clean up of pseudo deleted index entries as part of Db2 system task cleanup
- Clean up of XML multi-version
- Clean up of pseudo deleted index entries
- Portions of XML multi version documents cleanup
- System related asynchronous SRB processing with the exception of P-lock negotiation processing.
- **MSTR System related asynchronous SRB processing**
    - **Log write and log read**
- More Utilities eligibility

Db2 Version 12 Even More zIIP Exploitation

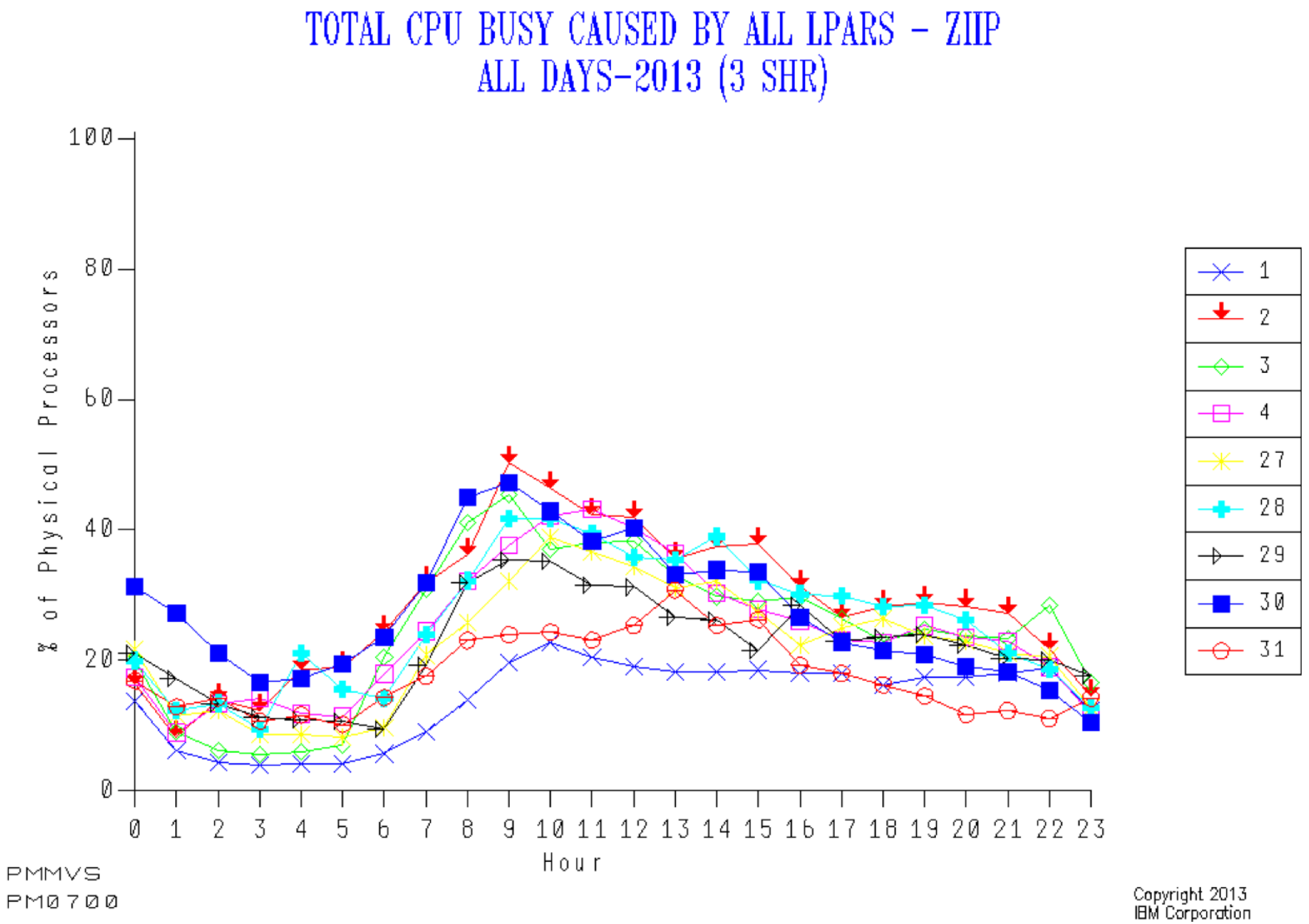
- Parallel child tasks of an SQL query can now be 100% zIIP eligible
  - Previously only 80%
- REORG: Reload phase is now 100% zIIP eligible
  - ~up to 17% more for REORG
- LOAD: Reload phase is now 100% zIIP eligible
  - ~up to 90% more for LOAD
  - PI73882 initial Non-Parallel implementation
  - PI80243 added Parallel
- Automatic GRECP/LPL recovery retry is 100% zIIP eligible
- Task which checks every 2 minutes whether an index is an FTB candidate
  - Fast Traverse Blocks (FTBs) are in-memory structures for non-leaf index pages


If You Only Remember 4 Things
The Importance Of zIIP Capacity Planning
Recentish News
Db2 Criticality
So Let's Take This Seriously...
Performance
Miscellaneous Topics
Conclusion
Appendix - Further Db2 Information

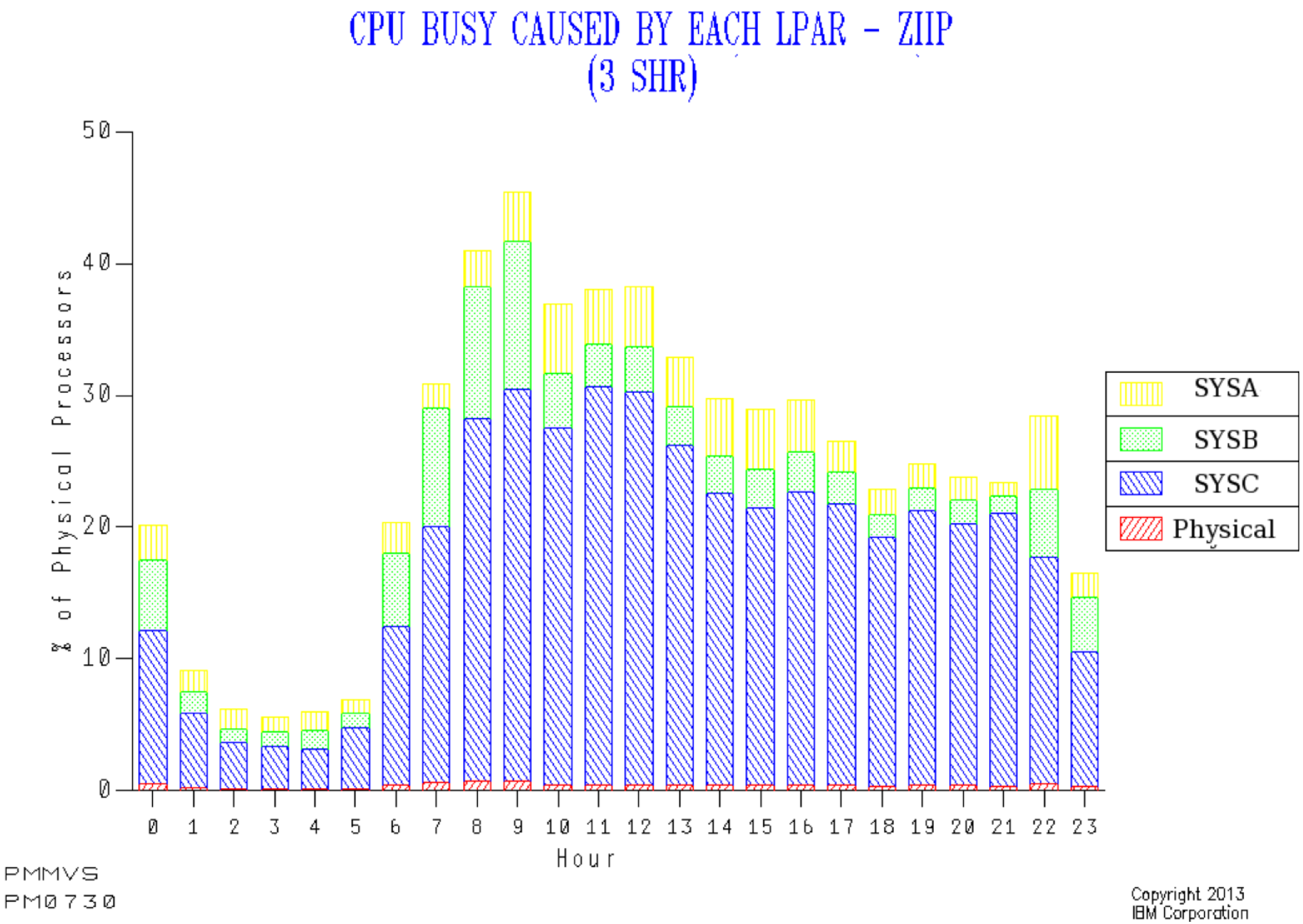
Many Sources Of Information

- From RMF:
  - SMF 70
  - SMF 72-3
- From z/OS:
  - SMF 30 at the address space level
  - SMF 113, 99-14 at the processor level
- From middleware:
  - SMF 101 Db2 Accounting Trace
  - SMF 110 CICS Statistics Trace
- **NOTE:** SMT adds considerable metrics complexity

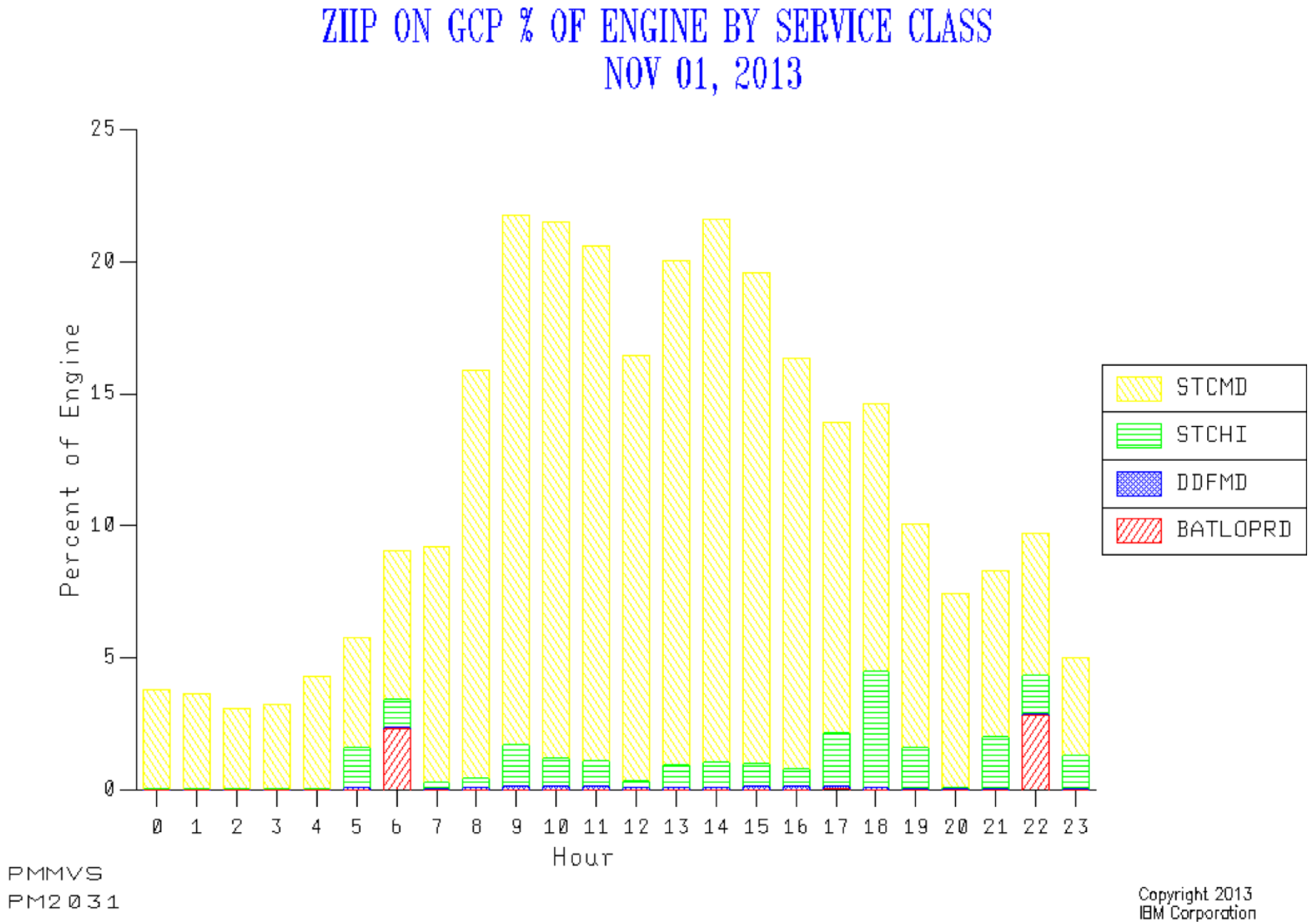
From 70-1 You Can See How Busy The zIIP Pool Is



You Can Also See Which LPARs Use zIIP The Most



From 72-3 You Can Break zIIP Usage Down By Service Class Or Report Class (Slide is for zIIP on GCP)



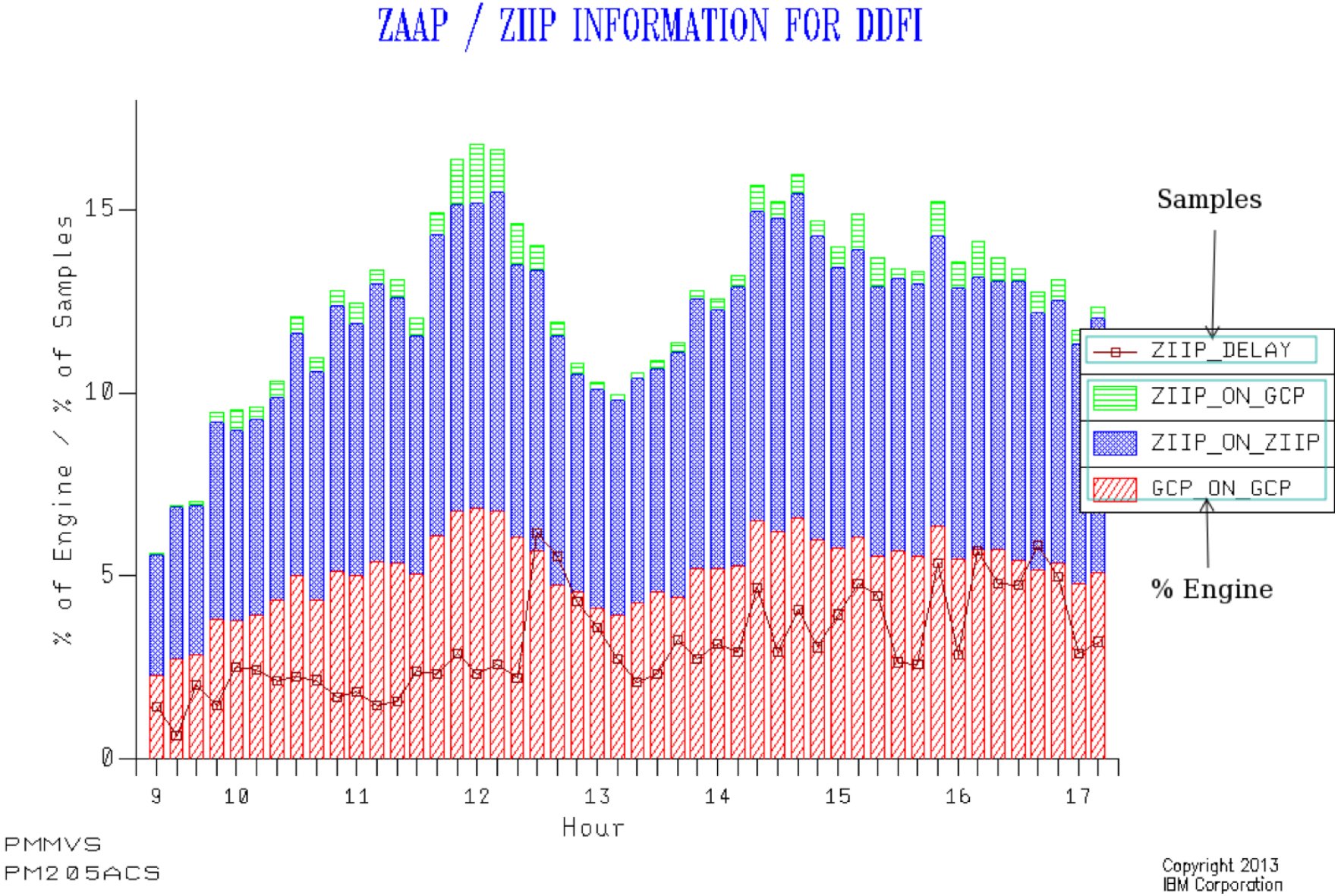
Velocity Calculation

$$\text{Velocity} = \frac{\text{Using}}{\text{Using} + \text{Delay}} \times 100 \quad (\%)$$

Service Class (Period) Information

- For Velocity Goal service class periods you get samples...
    - For work that's not zIIP eligible you get e.g. Using and Delay samples for CPU and I/O
        - These are the sample counters that affect the velocity calculation
- With zIIP-eligible work you get better resolution – as you get Using and Delay samples specific to zIIP
- Useful to correlate Delay for zIIP with % zIIP on GCP
    - Can explain behaviour
    - **Note:** Zero zIIP on GCP is not necessarily a sign all is well

A Sample DDF-Related Service Class



Understanding Your Exploiters

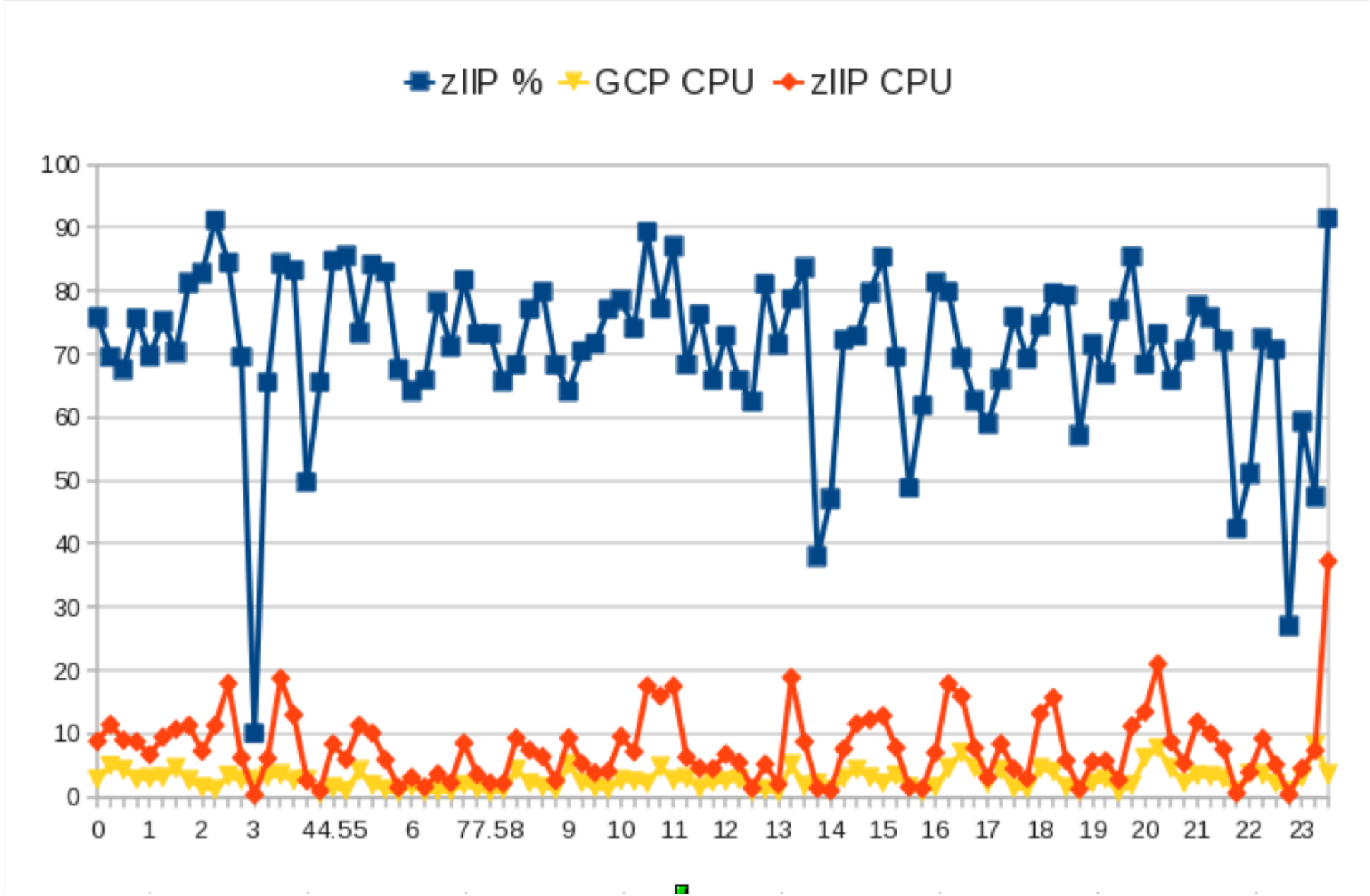
- Exploiters come in different shapes and sizes
    - And leave “footprints” accordingly
    - Likewise different criticality levels
        - Which **ought** to be reflected in WLM setup
- Dependent Enclave
    - e.g. DBM1 Prefetch / Deferred Write
    - Runs in address space's Service Class
- Independent Enclave
    - e.g. DDF
    - Runs in its own Service Class
- zAAP-on-zIIP
    - Formerly zAAP-eligible e.g. Java
- e.g. Dependent Enclave for DBM1 Address Space is Performance Critical from Db2 Version 10

SMF 30 Shift-Level Summary Of zIIP Exploiters. This Is Db2 V10 - V11 Would Show MSTR

Job Name	Program	TCB %	Dependent Enclave %	Independent Enclave %	zIIP on GCP %	zIIP on GCP - Dependent Enclave %	zIIP on GCP - Independent Enclave %	zIIP on GCP - Other %	zIIP %
Workload: STC Service Class: STCHI Report Class: STCDB2									
DSNRDBM1	DSNYASCP	1.0	0.8		0.84	All			
DSNRDIST	DSNYASCP	0.3		0.2	0.14		All		
Workload: STC Service Class: STCMD Report Class: STCCOTH									
CTGSTRA1	CTGBATCH	0.2			0.12			All	
CTGSTRA3	CTGBATCH	0.2			0.12			All	
CTGSTRA4	CTGBATCH	0.2			0.12			All	
CTGSTRB1	CTGBATCH	0.2			0.12			All	
CTGSTRB3	CTGBATCH	0.2			0.12			All	
CTGSTRF1	CTGBATCH	0.8			0.65			All	
CTGSTRF3	CTGBATCH	0.8			0.65			All	
CTGSTRG1	CTGBATCH	0.8			0.65			All	
CTGSTRG3	CTGBATCH	0.8			0.65			All	

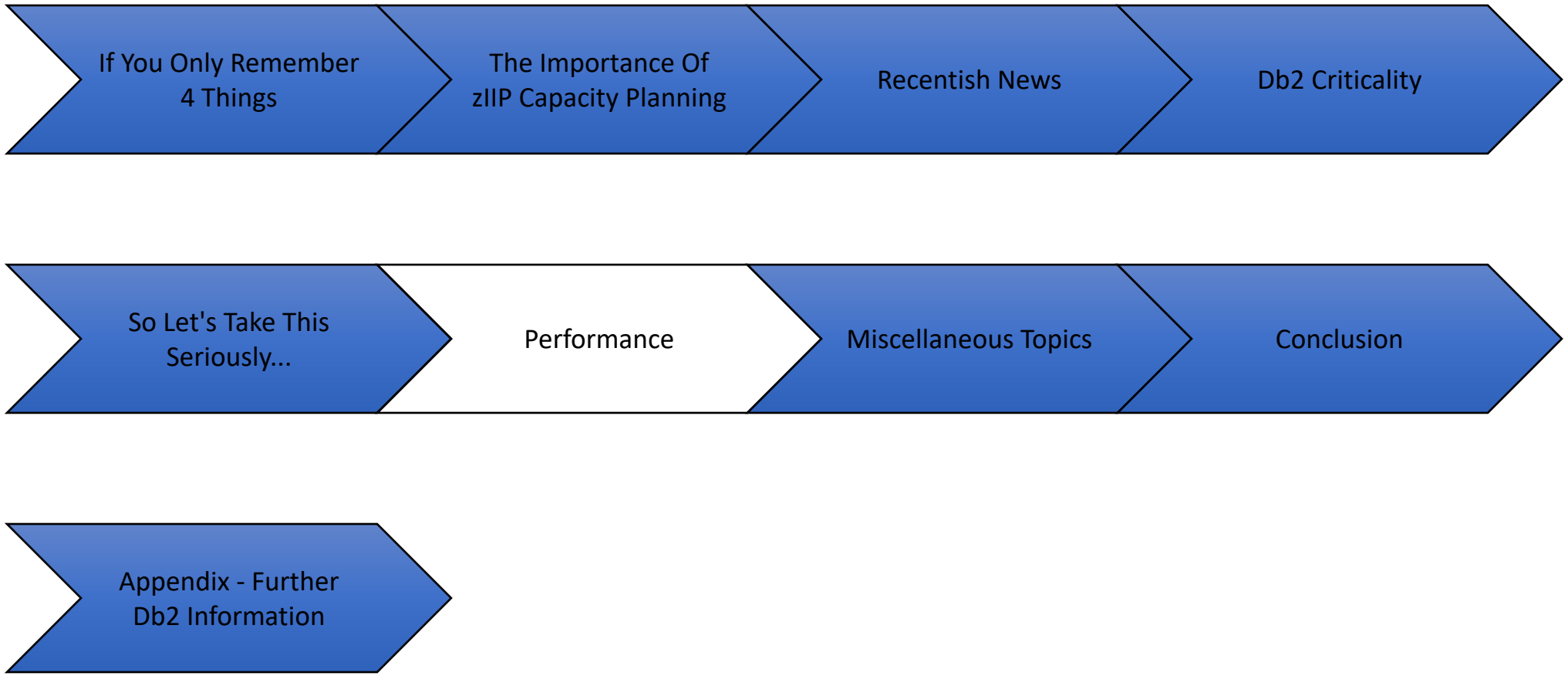
A Db2 Version 10 DBM1 Address Space's zIIP and CPU Usage Over 24 Hours

Note: In Versions 11 and 12 an even higher proportion is zIIP exploitative



Projections

- Projecting Near-Term zIIP CPU Requirements:
  - For existing workloads that don't run on a zIIP:
    - “Project CPU” reports how much CPU is zIIP-eligible via RMF
    - Db2 V10 to V11 or later:
      - The remainder becoming zIIP-eligible is a reasonable estimate
        - Probably about 20%
        - None of it any less “stringent”
        - MSTR CPU has some zIIP eligibility
  - For workloads that don't exist yet use any sizing guidelines you can get
  - Some “adaptive exploiters” exist:
    - These only try to use zIIP if one configured
  - “Project CPU” will show zero
- Projecting Into The Future
  - The same as standard CPU Capacity Planning

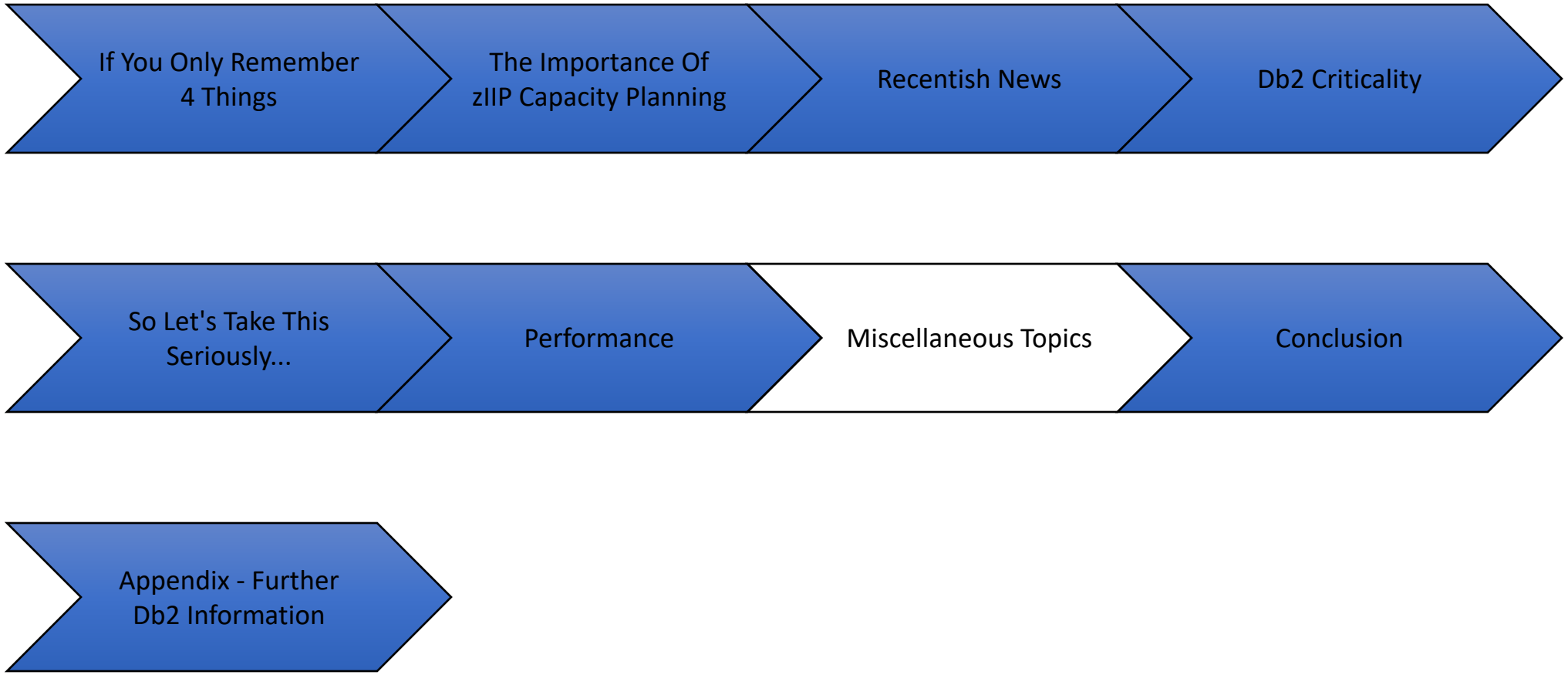

If You Only Remember 4 Things
The Importance Of zIIP Capacity Planning
Recentish News
Db2 Criticality
So Let's Take This Seriously...
Performance
Miscellaneous Topics
Conclusion
Appendix - Further Db2 Information

Care Needed Over LPAR Configuration

- An LPAR with low zIIP usage might not warrant any zIIP capacity
  - A trade off between lost zIIP exploitation and performance
  - Other LPARs might warrant better access to the zIIP more than this LPAR does
- Weights need especial care
  - One Vertical High (VH)in the zIIP processor pool guarantees "needs help" will work in a timely manner
    - There can be PR/SM delays when there is no VH, which can impact "needs help"
    - Vertical Medium (VM) or Low (VL) processors might not invoke "needs help" so soon
      - So other processors might not help run the zIIP work soon enough
      - Especially important where Db2 DBM1 and MSTR are the predominant zIIP users
      - So consider carefully whether an LPAR that can't have a VH should have a logical zIIP
  - Weights are a "zero sum" game
    - LPAR A having too high a weight prevents LPAR B from having a high enough share
- zIIPs dispatched in eg another drawer can affect processor cache effectiveness
  - Running on VM's and, especially, VL's increases this risk

zIIP “Short Engine” Effect

- LPARs with low zIIP usage sharing a zIIP
  - Especially with a low weight
- Potential high latency in logical zIIP being dispatched
  - If other LPARs have the zIIP
- In multiple zIIP case Hiperdispatch can help
  - “Corrals” work into fewer logical (& physical) zIIPs
  - But e.g. 1-Way Queuing regime has unhelpful characteristics
  - zIIP Parking rarely seen
- Of especial concern:
  - Single zIIP LPAR with low weight and only “DBM1 Engines” exploiter
  - Consider not configuring the zIIP to the LPAR


If You Only Remember 4 Things
The Importance Of zIIP Capacity Planning
Recentish News
Db2 Criticality
So Let's Take This Seriously...
Performance
Miscellaneous Topics
Conclusion
Appendix - Further Db2 Information

Service-Class Level IIPHONORPRIORITY

- Introduced with OA50845 / OA50953
- Prevents “needs help” processing by GCPs
    - Meaning: Prevents ”cross over”
- Designed to protect the GCPs from large zIIP exploiters such as Spark
    - Spark is highly zIIP-eligible
    - Can use a lot of CPU
    - The Service Class can also be memory capped
- SMF 30 has a flag for this
    - At the address space level
- Could, in principle, be applied to non-Spark work
- Could be a useful tool

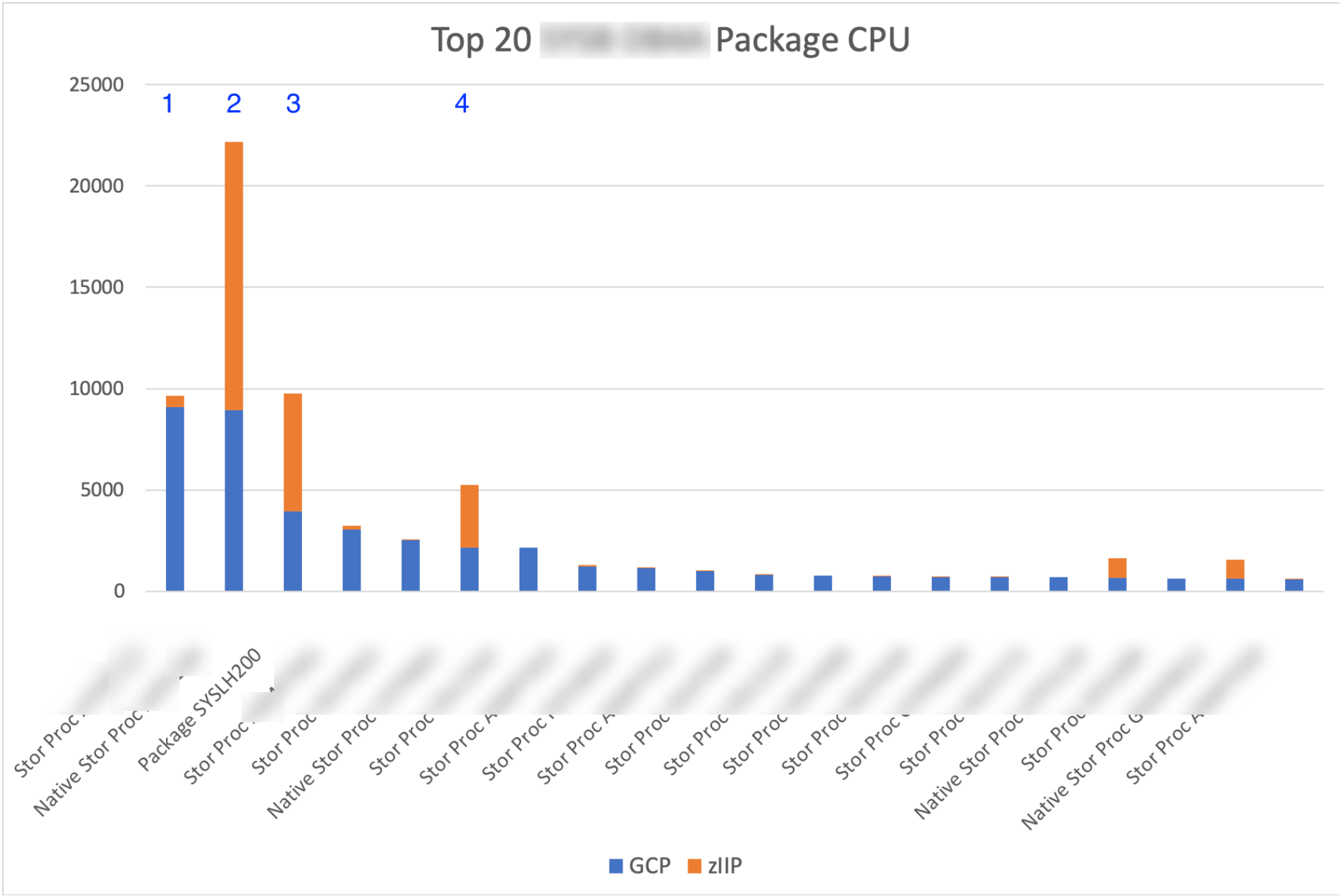
zCX Container Extensions

- Available with z/OS 2.4 and requires z14 or later hardware
- Allows Docker containers for Z to run on z/OS
  - About 3000 Docker images available for Z as of January 2020
- Contains enough Linux and Docker infrastructure to run the containers
- Can be multiple containers (application instances) per zCX address space
- Can be multiple zCX address spaces per z/OS LPAR
- **High level of zIIP eligibility**
  - You'd better know if they arrive and measure their impact
  - Might be a candidate for Service-Class Level IIPHONORPRIORITY
- Might drive Db2 zIIP Usage
  - DDF independent enclaves and DBM1 address space
- You define how many virtual CPUs each zCX address space has
  - These map to TCBs but are seen by Docker as processors
  - Controls ability to access zIIP (and GCP)
- zCX currently (1st release) uses 4K Preferred fixed pages to back Linux Guest's storage
  - These can be large (2GB minimum) so you also need to make sure you do proper memory capacity planning

zIIP And Db2 Accounting Trace

- zIIP Eligibility complex question for Db2 work
- Db2 Accounting Trace useful here
    - With Trace Classes 1,2,3:
        - QWACCLS1_zIIP - Class 1 zIIP Time
        - QWACCLS2_zIIP - Class 2 zIIP Time
        - QWACTRTT_zIIP - Trigger zIIP Time
        - QWACSPNF_zIIP - Stored Procedure zIIP Time
        - QWACUDFNF_zIIP DS CL8 - User Defined Function (UDF) zIIP Time
        - QWACZIIP_ELIGIBLE DS - zIIP-on-GCP Time
    - With Trace Classes 7,8:
        - QPACCLS7_zIIP - Package-level zIIP Time

zIIP Eligibility For Db2 Stored Procedures Depends On Whether Native Or Not



zIIP To GCP Ratios

- What's most important is how the zIIP performs
    - Especially how CPU-stringent workloads perform
- Extreme case 12 : 1 GCP : zIIP ratio especially unhelpful because of the “1”
- An unscientific view is that somewhere in the region of 2 : 1 to 4 : 1 is rarely harmful
- Consider both “in LPAR” and “across the machine” effects
    - Often the answer will be “no zIIPs for this LPAR”
        - Especially when the 1 zIIP is shared and not a VH

Subcapacity General Purpose Processors

- zIIP runs at full speed
    - e.g. about 65-75% faster than a 6xx General Purpose Processor (GCP)
- Good news:
    - zIIP higher capacity and faster than GCP
- Not so good news:
    - zIIP-eligible work running on a GCP is worse
    - Variability in speed if some runs on zIIP and some not
        - Not terribly different to DDF situation since PM12256
- How this plays depends on numbers of threads and CPU-intensiveness
- CPU Conversion Factor:
    - Use R723NFFS to convert zIIP CPU to GCP equivalent (and vice versa)
- Consider configuring more zIIPs than raw capacity requirement suggests
- SMT for zIIPs might counteract slower GCPs to some extent

Simultaneous Multithreading (SMT)

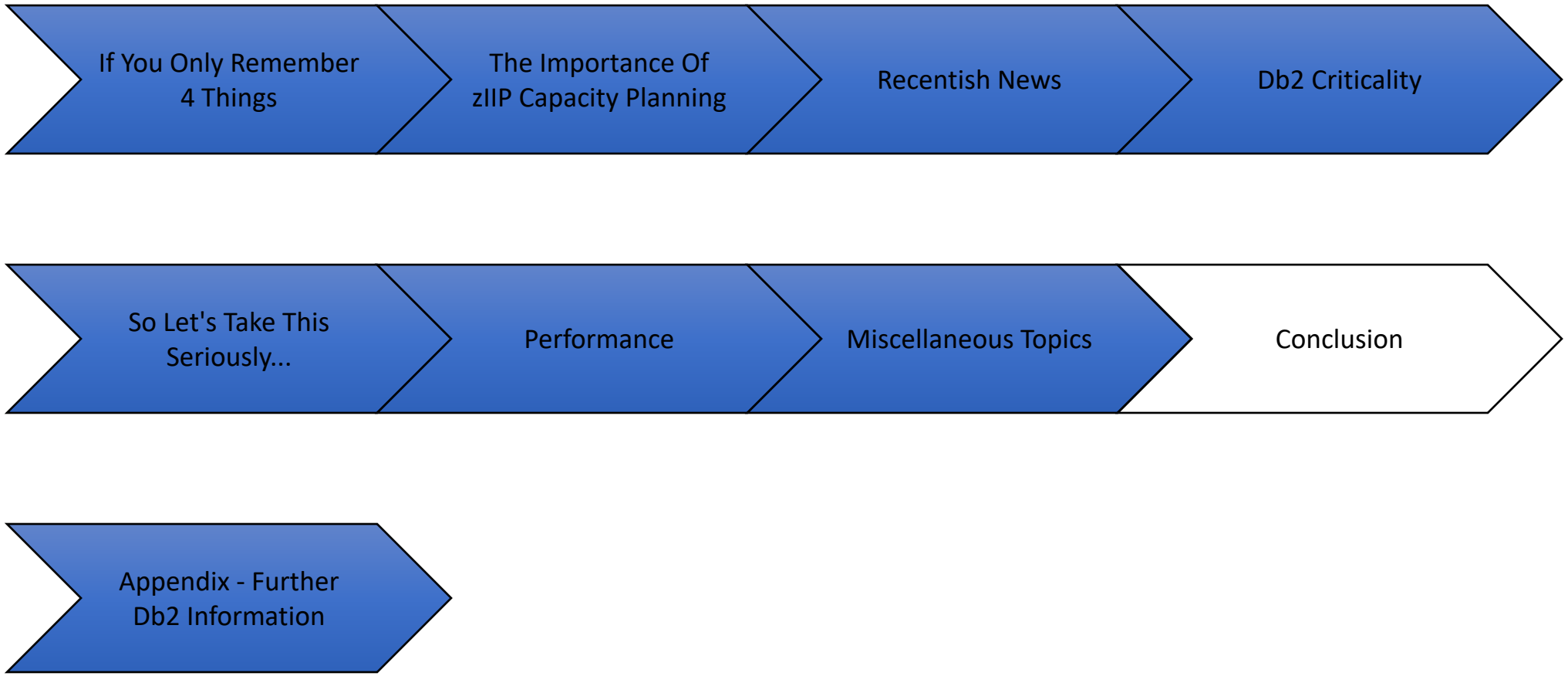
- Available with z13 for zIIPs (and IFLs)
    - (Mandatory with z14 for SAPs/IOPs)
- Allows two threads or CPUs on the same core
- On or off at LPAR level
- Consider the case of an SMT ratio of 1.25
    - Implies each thread runs at 0.625 times the speed of non-SMT
- Good for Db2 DBM1 "engines"
    - Not so good for monolithic java batch jobs
        - Though might alleviate queuing
- Instrumentation **necessarily** complex

z15 System Recovery Boost (SRB)

- Can allow zIIPs to run non-zIIP-eligible shutdown and restart work
  - During a specified boost period
  - With no impact on IBM software cost
  - Also allows subcapacity GCPs to run at full speed
- Using SRB Upgrade allows deployment of extra zIIPs
- SRB is on by default for z/OS
  - For some LPARs you might disable it - eg for test LPARs where SRB isn't part of the test
- During a boost period an LPAR could use a lot of zIIP
  - To another LPAR's detriment if the other LPAR's share doesnt protect it
  - Of course, this could be true of **any** large surge in zIIP-eligible work

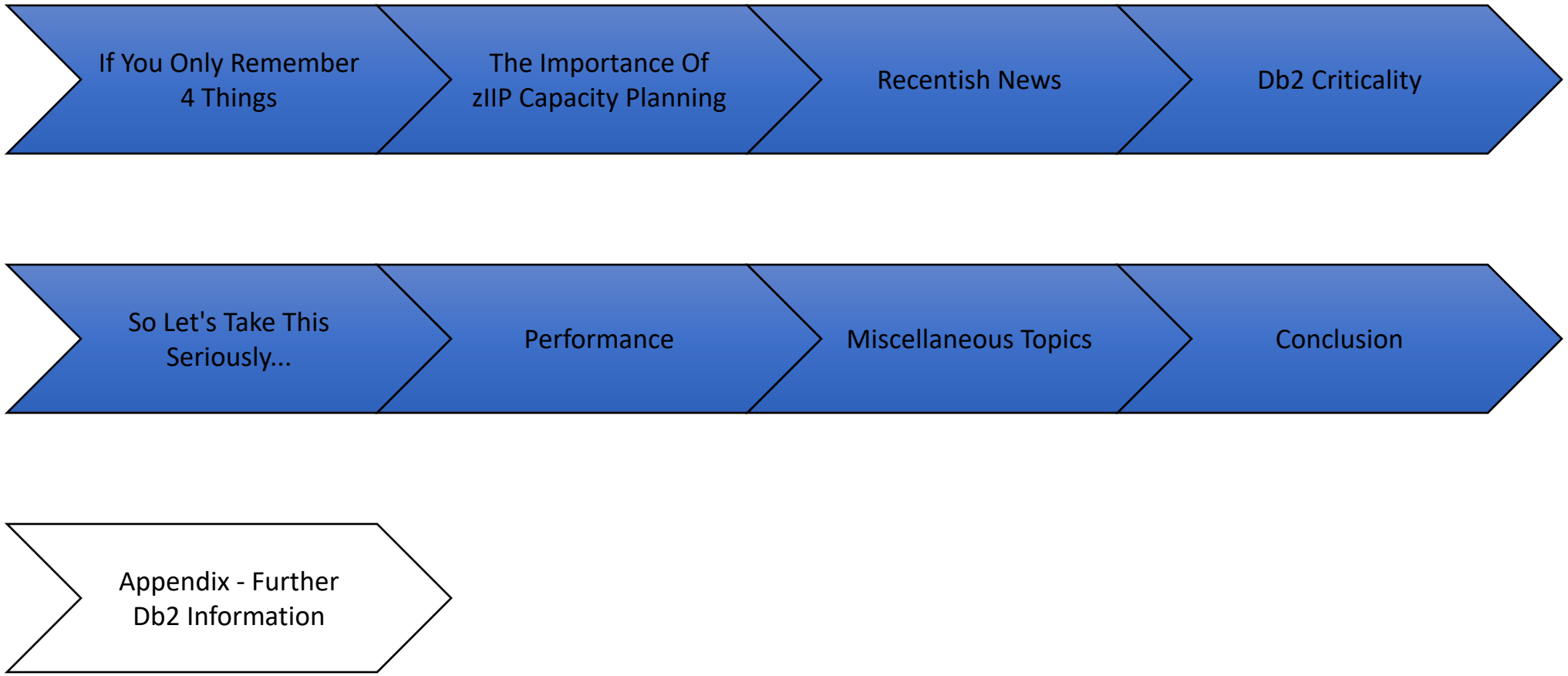
z15 System Recovery Boost (SRB) ...

- Assess how you choreograph shut down and restart of an LPAR
    - To potentially increase parallelism and reduce elapsed times
    - eg with SMF 30 for relevant address spaces
    - Also your automation for shutting down and restarting
        - Possibly around SRB's messages
        - Use z/OS IEASDBS proc to activate shutdown boost
- RMF support in APAR OA56683
    - RMF Product Section for all 7x record types tells you if this LPAR has SRB on and which kind
    - In SMF 70-1 Partition Data Section tells you the same for ALL LPARs on the machine


If You Only Remember 4 Things
The Importance Of zIIP Capacity Planning
Recentish News
Db2 Criticality
So Let's Take This Seriously...
Performance
Miscellaneous Topics
Conclusion
Appendix - Further Db2 Information

If You Only Remembered 4 Things I Hope They Were

- Capacity Planning for zIIPs has become much more important
- Pay especial attention to what % Busy represents “full”
    - It's probably much lower than for GCPs
- Measure zIIP potential and usage down to the address space level
- Consider LPAR configuration carefully for zIIPs


If You Only Remember 4 Things
The Importance Of zIIP Capacity Planning
Recentish News
Db2 Criticality
So Let's Take This Seriously...
Performance
Miscellaneous Topics
Conclusion
Appendix - Further Db2 Information

Why Protect Db2 System Address Spaces From Being Pre-empted? 1 / 2 (Thanks To John Campbell)

- **MSTR** contains the Db2 system monitor task
  - Monitors CPU stalls & virtual storage constraints
- **DBM1** manages Db2 threads & critical for local Db2 latch & cross-system locking negotiation
  - Delays in negotiating a critical system or application resource can lead to a slowdown of the whole Db2 data sharing group
    - Example: P-lock on a space map page

Why Protect Db2 System Address Spaces From Being Pre-empted? 2 / 2

- **DIST** & **WLM-Managed Stored Procedure Server AS** only run the Db2 service tasks
  - i.e. work performed for Db2 not attributable to a single user
  - Classification of incoming workload, scheduling of external stored procedures, etc.
  - Typically means these address spaces place a minimal CPU load on the system
    - **But** they do require minimal CPU delay to ensure good system wide performance
  - The higher CPU demands to run DDF and/or SP workloads are controlled by the WLM service class definitions for the DDF enclave workloads or the other workloads calling the SP
    - Clear separation between Db2 services which are long-running started tasks used to manage the environment and transaction workloads that run and use Db2 services

Further Information On MSTR zIIP starvation (Thanks To Adrian Burke)

- **Implications:**
  - I/O delays during Db2 log write activity (synchronous) cause:
    - Delays during 2-phase commit processing, GBP write activity, and Db2 index page split activity, which in turn causes
      - Delays in P-lock, index latch, and page latch processing which in turn causes
      - Transaction delays and transaction locks to be held longer, resulting in
      - Increased lock contention, timeouts, TMON thread cancels, and build-up of workload in Db2 and CICS
    - These events can cause system wide impacts as P-locks are global in nature preventing any other transaction from modifying that object
- **Investigation:**
  - SMF 42.6 maximum data set I/O response times
  - Db2 Class 3 Log Write I/O Delay and Update Commit times