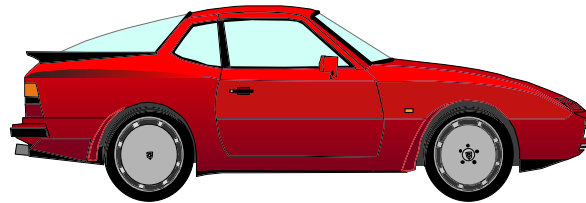


Pop the Hood on Workload Manager !



Steve Grabarits

STEVEJG@US.IBM.COM

S/390 Performance, Poughkeepsie, New York

Trademarks

The following terms used in this publication are trademarks of the IBM Corporation in the United States and/or other countries:

ACF/VTAM	MVS
CICS	MVS/ESA
DFSMS/MVS	OS/390
Enterprise Systems Architecture/390	S/390*
ES/9000	System/390*
ESA/390	S/390*
IBM*	VTAM*
IMS	

This presentation has not been submitted to any formal IBM test.

It is distributed on an "as is" basis without any warranty either expressed or implied.

Abstract

How are OS/390 systems tuned by WLM in goal mode? Get a feel for what's inside of WLM by taking a quick look at the high-level internals, and then see how it operates. The speaker will explore a few case studies exhibiting different stress points, and will demonstrate how WLM managed the system. Some of the newer WLM functions in OS/390, such as discretionary goal management, will be discussed.

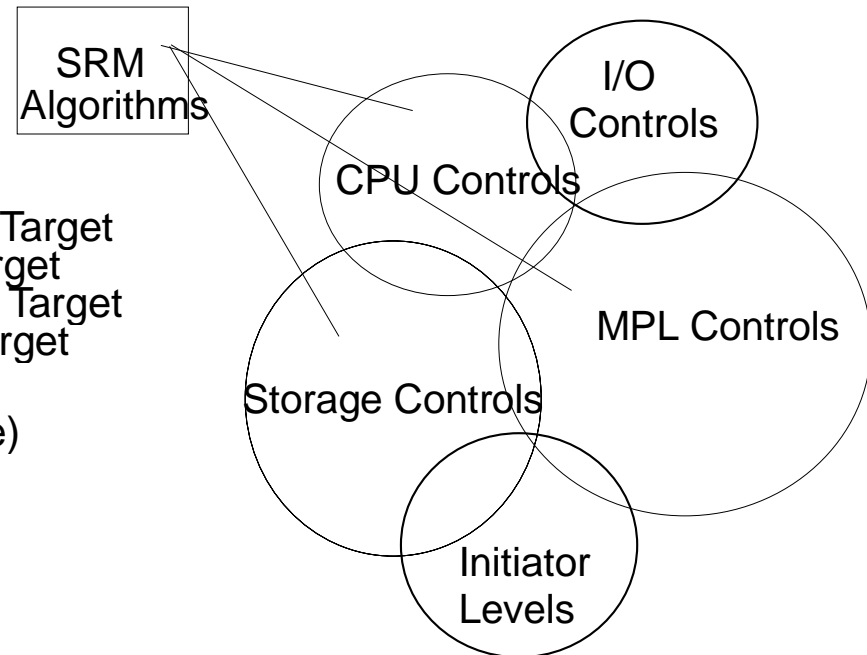
AGENDA

- WLM / SRM : Internals Overview
- Case Study
 - Multi-Workload including OLTP, CPU stress
- Recent Enhancements
 - I/O Priority Management
 - Batch Initiator Management
 - Discretionary Goal Management

Controls Managed by WLM

In goal mode, SRM uses the following controls to manage work to help meet customer specified goals.

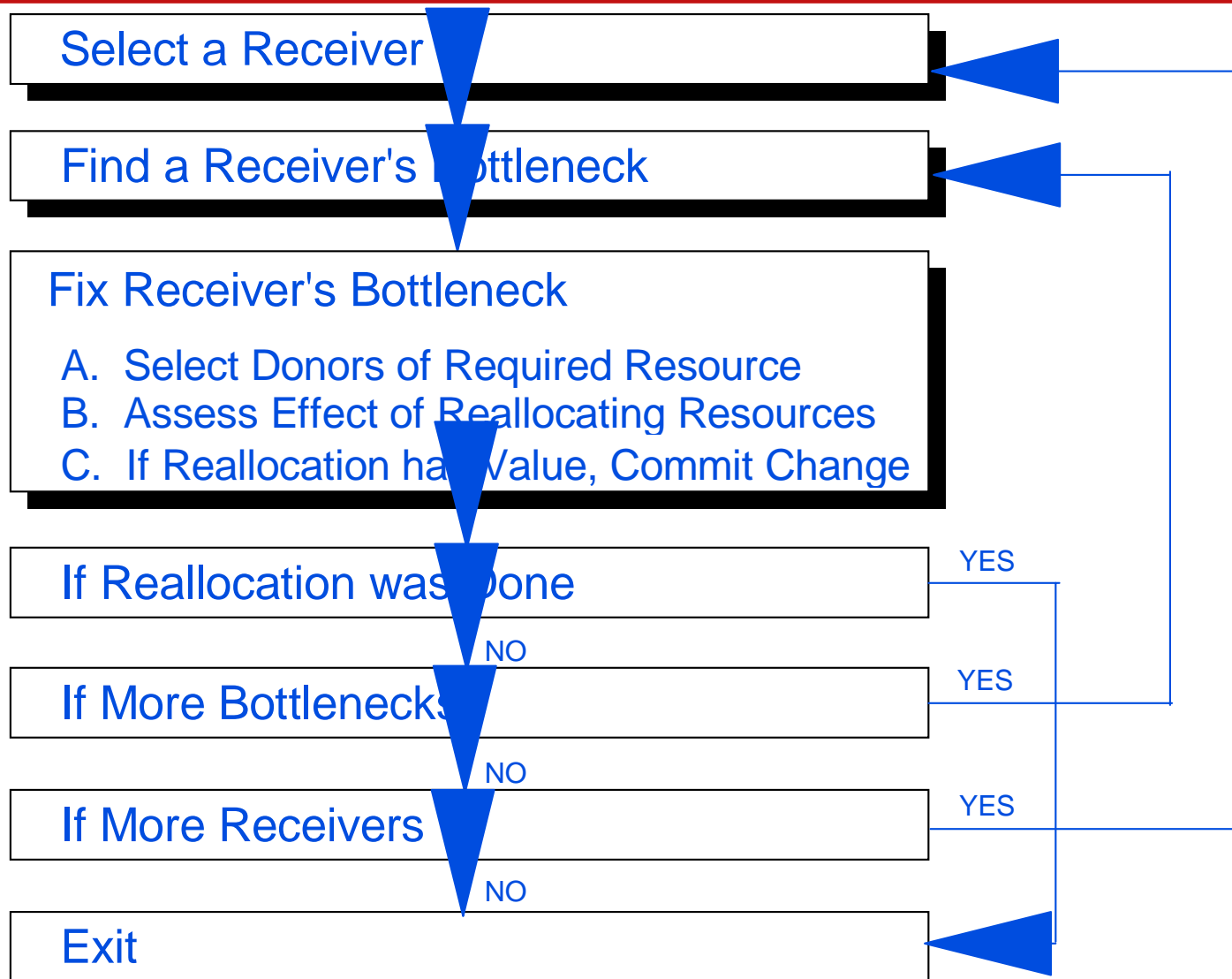
- ▶ Dispatching Priority
- ▶ MPL Targets
- ▶ Swap Protect Time
- ▶ Storage Targets
 - ▶ Protective Processor Storage Target
 - ▶ Protective Central Storage Target
 - ▶ Restrictive Processor Storage Target
 - ▶ Restrictive Central Storage Target
- ▶ Expanded Storage Policies (protected, LRU, space available)
- ▶ Swap working set pages
 - ▶ VIO pages
 - ▶ Hiperspace pages
 - ▶ Stolen pages and swap trim
- ▶ Resource Group Cap Slices
- ▶ I/O Priority
- ▶ Initiator Levels (Server Address Spaces)



WLM/SRM Philosophy

- ▶ Dynamic Workload Characterization
 - based on state sampling and measurements
- ▶ Prediction for Resource Adjustments
 - 'Rules of Thumb' are not used
 - Actual Measured Results: Plots and Histories
- ▶ Receiver-Donor Loop to adjust resources
 - every 10 seconds
- ▶ Sufficient for Adjustment to make an Improvement
 - Single Problem at a Time
 - Optimal change at any given point not required

Policy Adjustment Loop



Performance Index (aka PI)

- ▶ Indicator of how well the service class period goal is being achieved.
 - PI < 1 means that the goal is being exceeded.
 - PI = 1 means that the goal is exactly being met.
 - PI > 1 means that the goal is being missed.

$$\text{Average Response Time Goal PI} = \frac{\text{Actual Average Response Time}}{\text{Goal Average Response Time}}$$

$$\text{Execution Velocity PI} = \frac{\text{Goal Execution Velocity \%}}{\text{Actual Execution Velocity \%}}$$

$$\text{Percentile Response Time Goal PI} = \frac{\text{Actual}}{\text{Goal}}$$

- ◆ Where 'Actual' means the actual response time that was actually achieved for the percentage of the goal.

Select a RECEIVER

Which Class of work to help? (which Service Class Period?)

- ▶ By Importance
- ▶ Act upon worst PI
 - by sysplex PI, then by local PI
 - OW25542 changed order, more emphasis on local PI
- ▶ Only select periods with $PI > 0.9$
- ▶ Resource Groups : factor in below group service minimum

Find Bottleneck

What is the largest delay that can be addressed?

- ▶ Processor Delay
- ▶ Aux Paging Delay
 - private, common, cross-memory, Hiperspace
- ▶ MPL Delay
- ▶ Swap In Delay
- ▶ I/O Delay
- ▶ Initiator Delay (QMPL)
- ▶ Processor Cap Delay

If period is being served, find largest delay impacting a supporting server period.

Fix Receiver's Bottleneck

Select potential donors of resource

- ▶ Generally, reverse order of receiver selection
 - discretionary is treated as importance 6
- ▶ Holding Resource in need (i.e. higher dispatch priority)
- ▶ Resource Groups Minimums (and PI) are factors

Assess Effect of Change

- ▶ Project change in delay samples
 - histories, plots, calculations
- ▶ Project PI deltas for receiver(s) and donor(s)

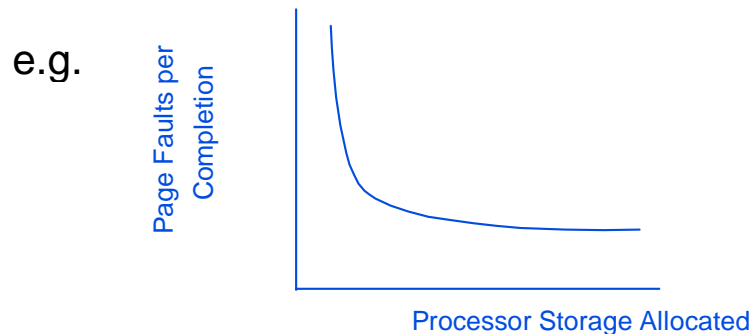
Projections

Dispatch Priority Actions

- ▶ Tables of processor characteristics of period
 - Maximum processor demand
 - Wait to Using Ratios
- ▶ State Machine: Combinations of moves evaluated

Storage Actions

- ▶ Use of Plots to track tuning relationships and delays



Value Checks

Net Value- Comparison of Receiver Benefit to Donor Loss

- ▶ If Receiver is more Important than Donor
 - Always make the move if Receiver is missing goal
- ▶ If Receiver is less Important than Donor
 - Never make the move if Donor is missing goal
 - or is projected to miss goal
- ▶ If Receiver and Donor are of equal importance
 - Receiver's PI benefit is more than Donor's loss and
 - Less disparity in projected PIs

Receiver Value - more than marginal improvement to Receiver

Resource Adjustment

- ▶ Run system efficiently such as
 - Shed work if system resources are overutilized
 - Bring in work if system resources are underutilized
 - Allocate central and expanded to minimize CPU cost of paging
- ▶ Not done at the expense of missing goals
- ▶ Includes and expands Working Set Management concepts
- ▶ Runs every 2 seconds

CPU Dispatch Priorities

255	SYSTEM
254	SYSTC
253	Small Consumer
252	<i>DYNAMIC PRIORITIES USED BY POLICY ADJUSTMENT</i>
208	
207	Unused
202	
201	Discretionary (MTTW)
192	
191	Quiesce

SMALL PROCESSOR CONSUMER

- periods with very little CPU
- move them out of way of critical adjustments

DYNAMIC PRIORITIES USED BY POLICY ADJUSTMENT

- periods with goals or server periods
- discretionary in a resource group with a min

Test Example

Multi-Workload including OLTP, high CPU Stress

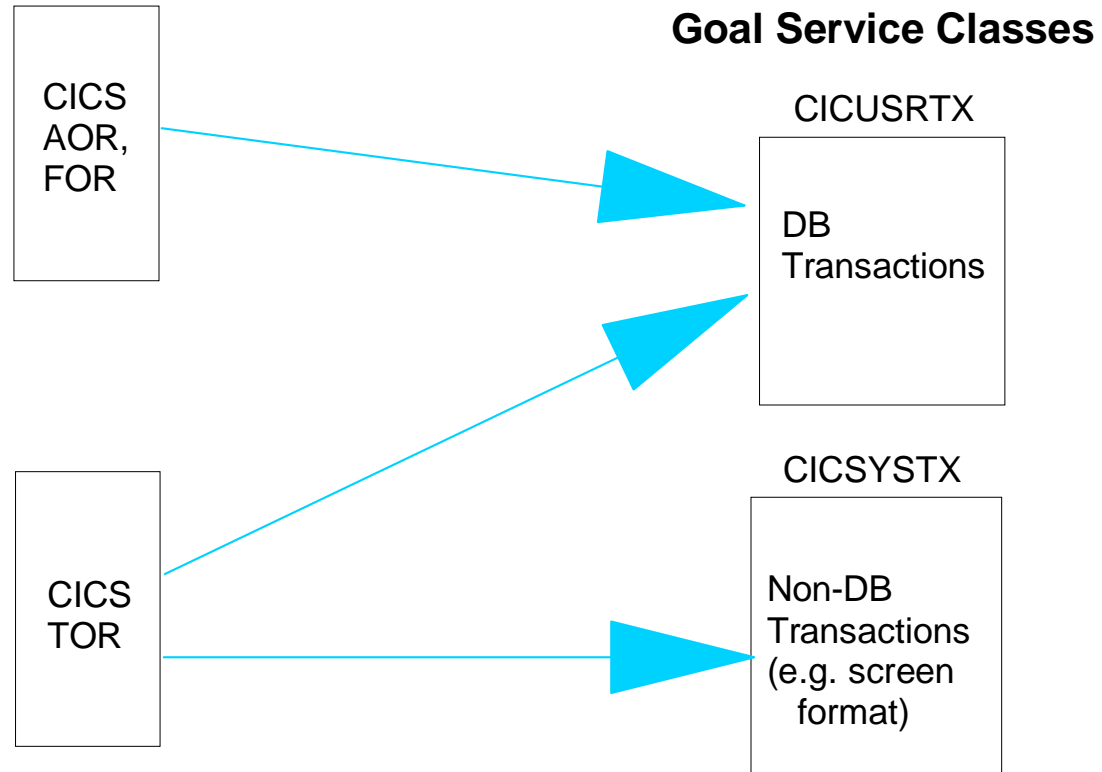
POLICY

Service Class	Goal Type	Value	Importance
CICUSRTX	Average Response Time	0.090 sec	High (2)
CICSYSTX	Average Response Time	0.015 sec	High (2)
IMS	Average Response Time	10.0 sec	Medium (3)
IMS1	Average Response Time	0.180 sec	Medium (3)
TSO (3 periods)	Avg Resp Time	0.100 sec	Medium (3)
	Avg Resp Time	1.000 sec	Medium (3)
	Avg Resp Time	3.000 sec	Low (4)
BATCHHI	Velocity	7 %	Lowest (5)
BATCHLO	Velocity	1 %	Lowest (5)

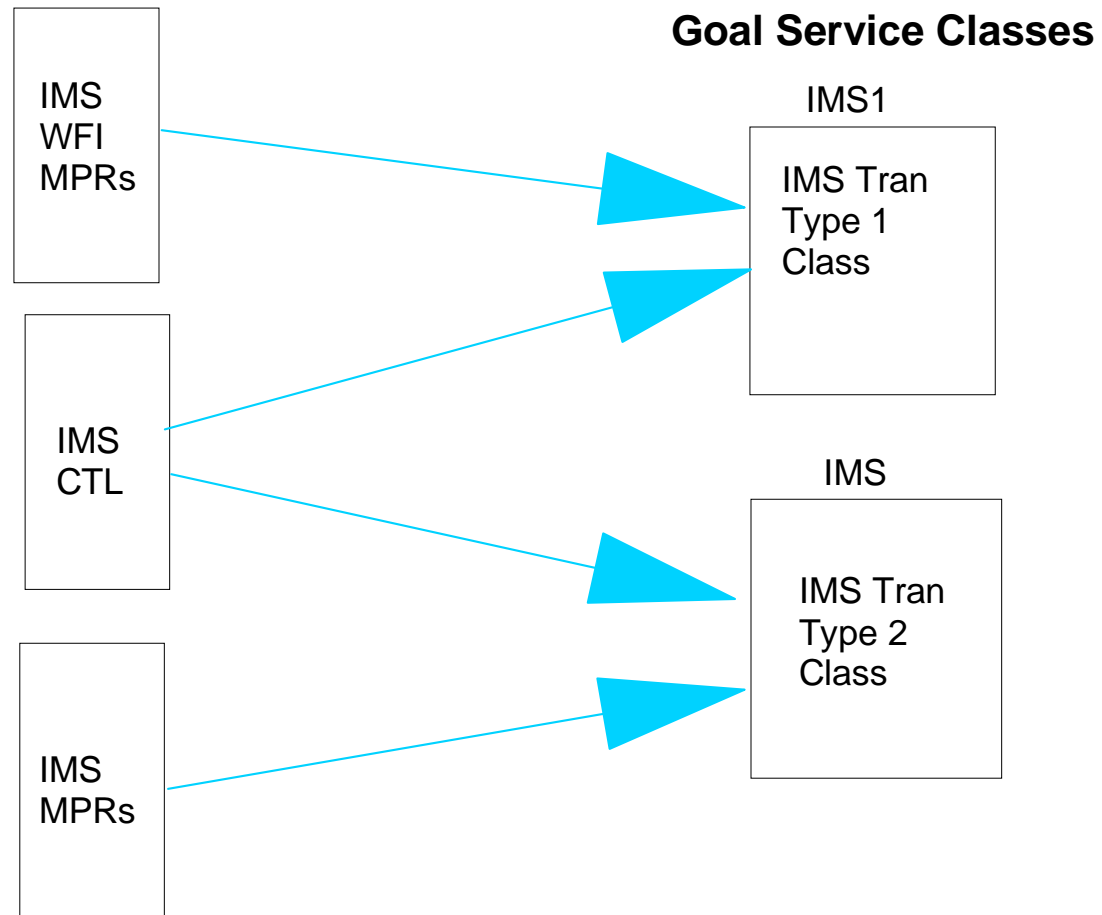
SYSSTC

VTAM
JES2
SMS
RMF
...

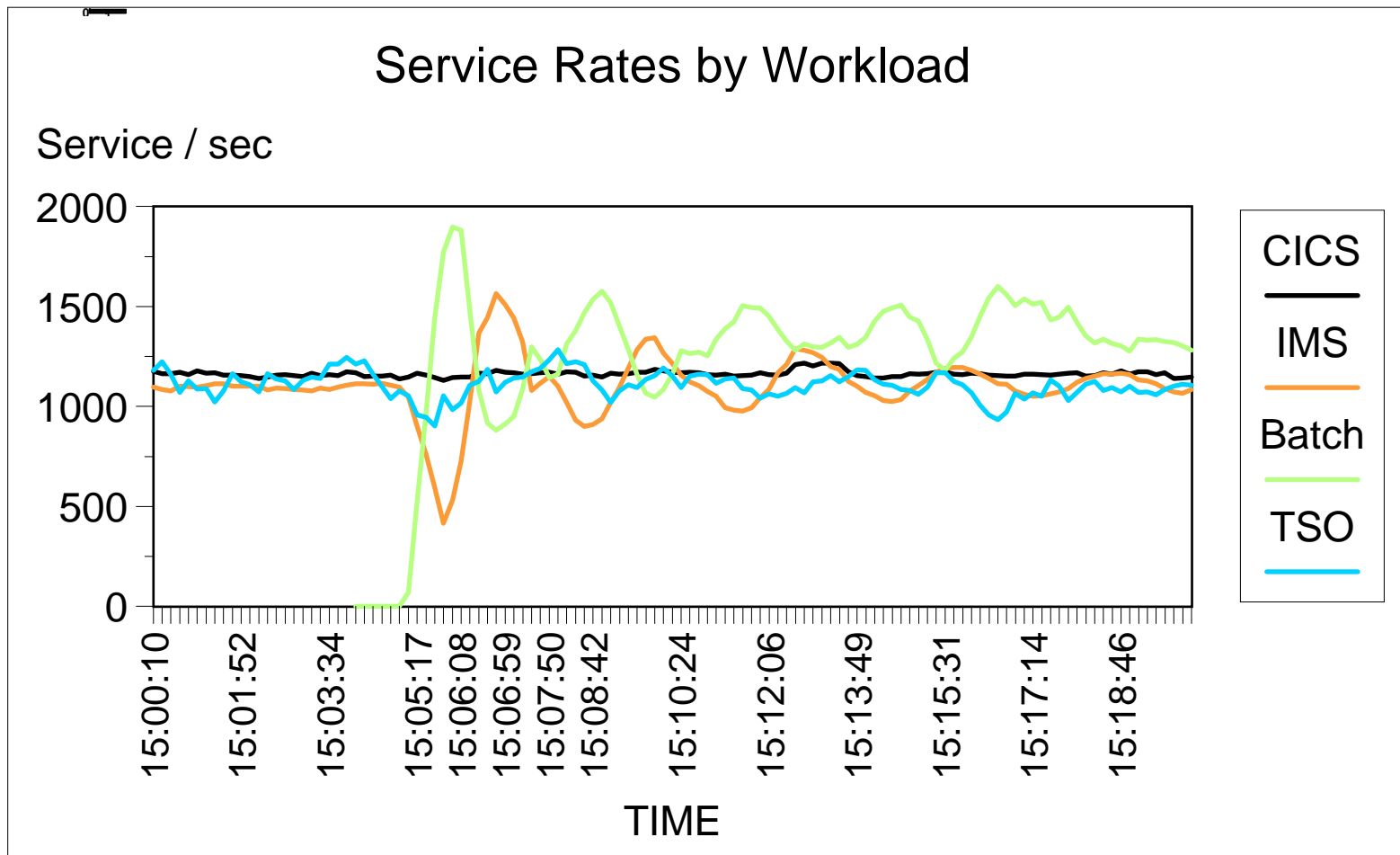
Transaction Servers



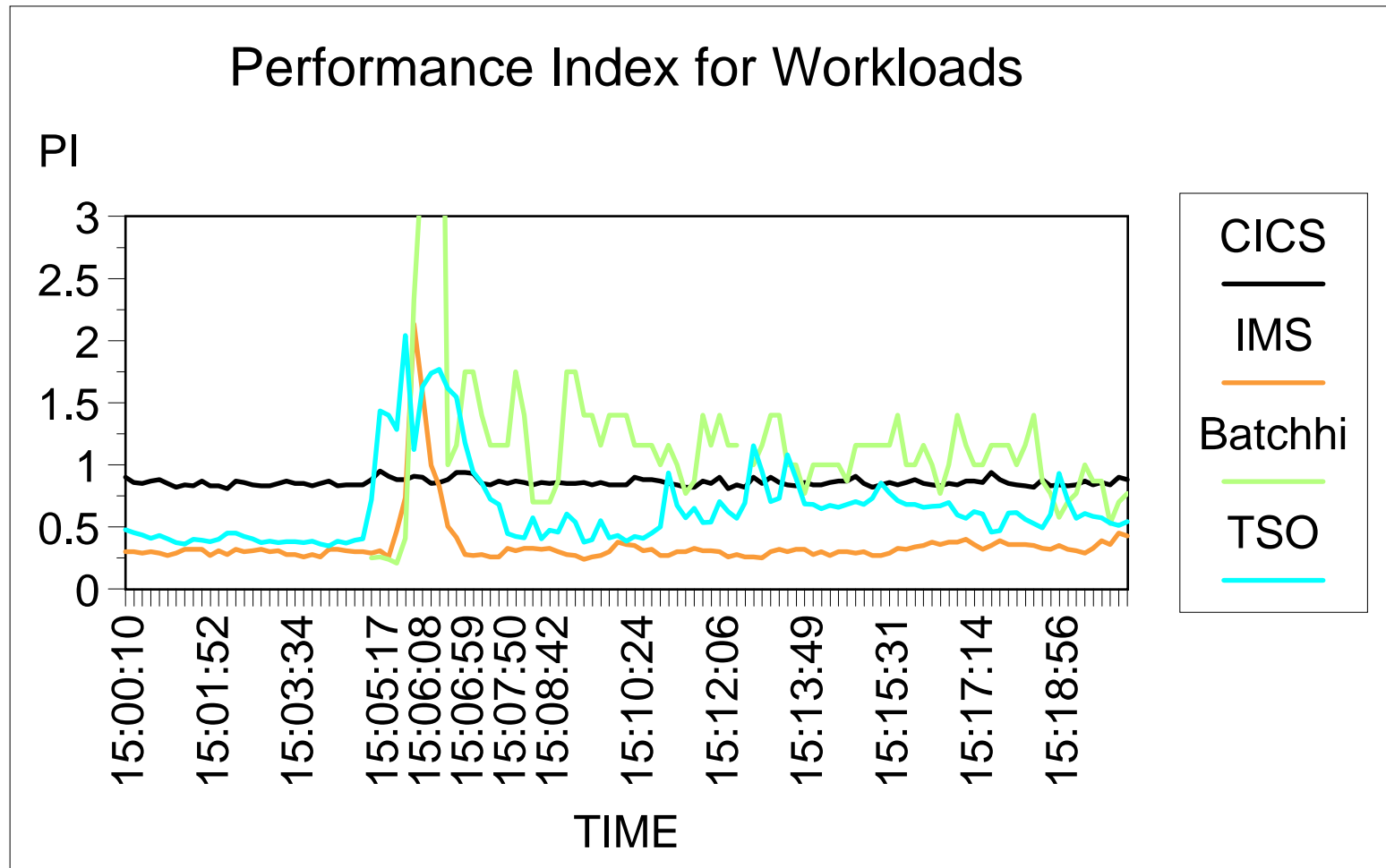
Transaction Servers (cont.)



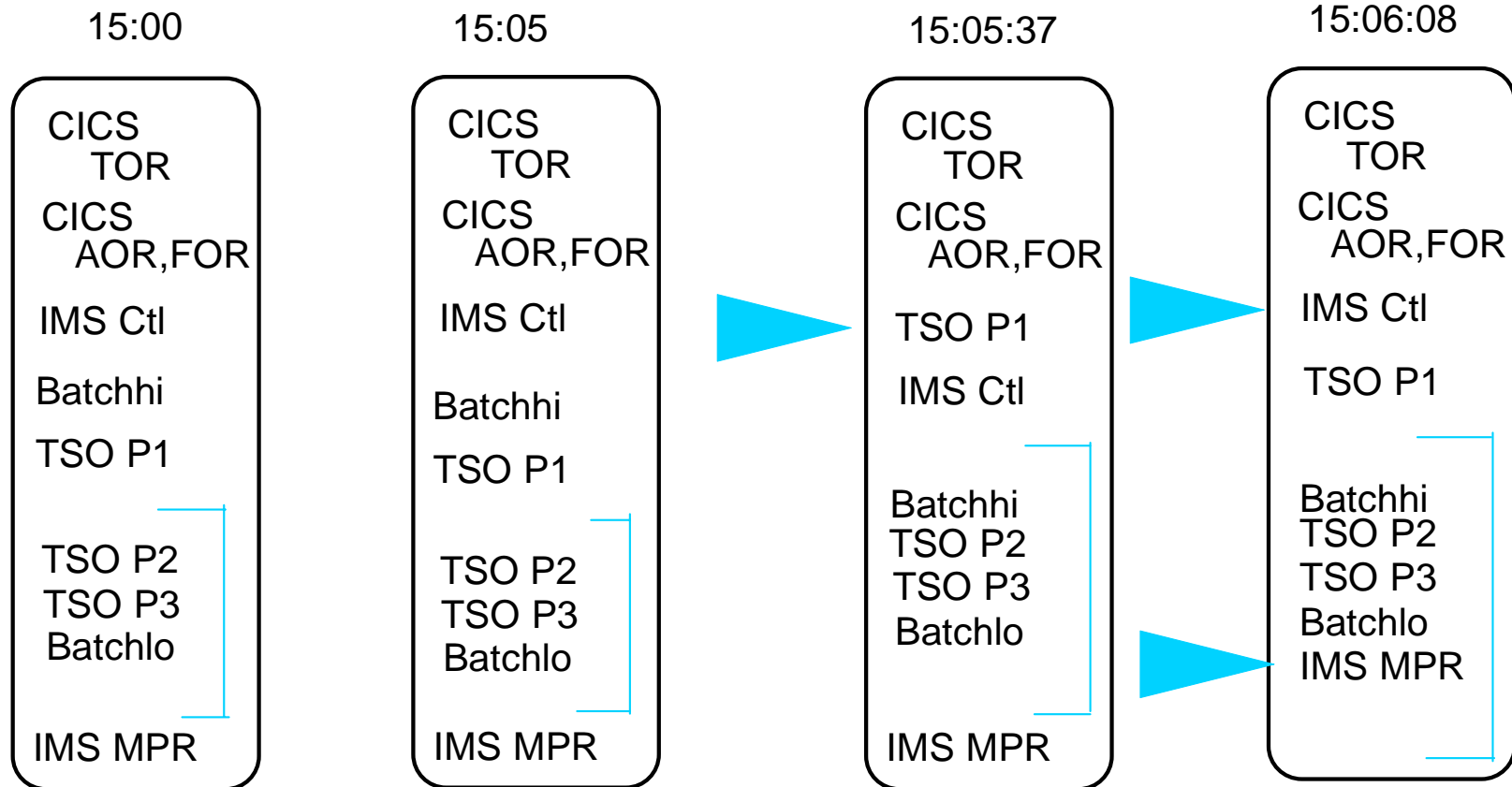
Service Distribution



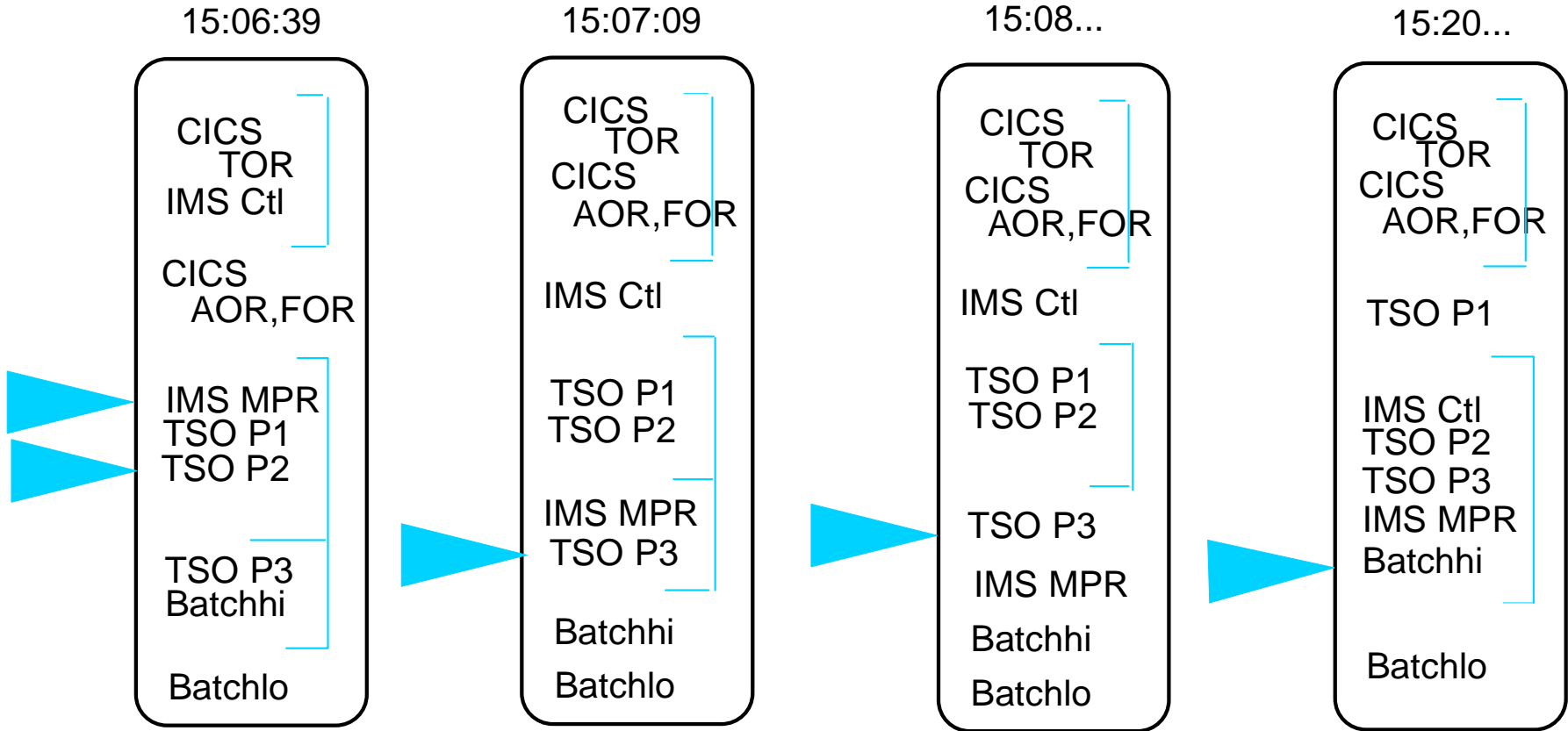
Performance Index



Dispatch Priority Adjustments



Dispatch Priority Adjustments (cont)



WLM/SRM Adjustments

- CPU adjustments made based on goals/importance
 - CICS most important: no effect
 - Batch least important: squeezed for CPU resource
 - Dynamic tradeoffs over time with IMS versus TSO
- Little storage contention: no significant storage actions
 - Transaction Server address spaces: protective central and processor storage targets

Recent Enhancements

OS/390 V1R3 I/O Priority Management

OS/390 V2R4 Batch Initiator Management

OS/390 V2R6 Discretionary Goal Management

I/O Priority Adjustments

- I/O becomes another resource for WLM to manage
- I/O samples (using + delay) included in execution velocity
- Policy adjustment code:
 - Find receiver's biggest bottleneck:
 - ▶ CPU delay
 - ▶ Storage delay
 - ▶ Now I/O Delay
 - Receiver's I/O delay can be address by:
 - ▶ Raising receiver's I/O priority or lowering a donor's I/O priority
 - ▶ Donor must be competing with receiver for at least some devices or action has no effect
 - ▶ Device Clustering used to determine this relationship

I/O Device Clustering

- Why device clustering?
 - When considering changing the priority of a service class period, WLM needs to be aware of the other periods competing for the same devices
 - These other periods might be affected by the priority change
 - Device Clustering is used to determine this relationship
 - Device Cluster - Set of service classes competing for the same or a subset of the same devices
- Characteristics of a device cluster are:
 - Each class is associated with a single device cluster
 - Each device used significantly is in a single device cluster
- Device clusters evaluated every 10 minutes

I/O Device Clustering Example

Service Class	Dev 200	Dev 201	Dev 202	Dev 500	Dev 501	Dev 502	Dev 503
Class 1	100	150	150	0	0	0	0
Class 2	0	90	100	0	0	0	0
Class 3	0	100	100	5	0	0	0
Class 4	0	0	0	100	100	100	100
Class 5	0	0	0	0	150	0	150

- Table shows how many times each service class sampled using or delayed for a set of devices
- Class 3's use of device 500 considered "insignificant"
- Clusters will be
 - Device Cluster 1: Class 1, 2, 3
 - Device Cluster 2: Class 4,5

I/O Priority Example (SMF 99)

10:40:32

CLASS	P	SPI	LPI	CODE	ACTION
TSOEVEN	2	37	37	8880	pa_io_rdon_cand
VEL50C	1	333	333	8620	pa_imuo_rec to 252
TSOEVEN	1	49	47	8960	pa_io_unc_sec_don
TSOEVEN	2	45	37	8940	pa_io_unc_don
VEL10J	1	76	66	8960	pa_io_unc_sec_don
VEL30D	1	214	176	8960	pa_io_unc_sec_don
VEL50C	1	178	333	8750	pa_io_inc_rec

Device Cluster Data

Cluster 6 Classes TSOEVEN VEL10J VEL30D VEL50C

I/O Priority Data

IOP	IMD	PMD	W2UR	Classes
252	86	546#	25	TSOEVEN 1 TSOEVEN 2
250	1438	978#	98	VEL10J 1 VEL30D 1 VEL50C 1
249	141	141	1312	TSOEVEN 3

Service Period Summary

Class	P	I	SPI	LPI	DP	CID	IP	State	Samples...
TSOEVEN	1	2	47	47	245	6	252	CPUU/472	IODU/48 OTHR/357 CPUD/204 IODD/60
TSOEVEN	2	2	37	37	245	6	252	CPUU/618	IODU/131 OTHR/582 IODD/569 AUXC/246
TSOEVEN	3	2	69	69	243	6	249	CPUU/152	IODU/48 IODD/249 OTHR/86 CPUD/70
VEL10J	1	2	66	66	245	6	250	CPUU/13	IODU/64 IODD/410 CPUD/9 OTHR/7
VEL30D	1	2	176	176	245	6	250	CPUU/22	IODU/95 IODD/506 CPUD/48 OTHR/11
VEL50C	1	2	333	333	251	6	250	CPUU/17	IODU/64 IODD/425 CPUD/9 OTHR/1
...									
VEL50E	1	2	250	250	247	4	250	CPUU/24	IODU/112 IODD/483 CPUD/37 OTHR/4
VEL50G	1	2	312	312	249	3	250	CPUU/19	IODU/65 IODD/420 OTHR/4
...									

WLM Batch Management in OS/390 V2R4

What WLM manages:

- Number of initiators
 - Queue delay samples identify delays due to lack of initiators
 - Queue delay included in velocity calculation
 - Adjustments based on goals and importance

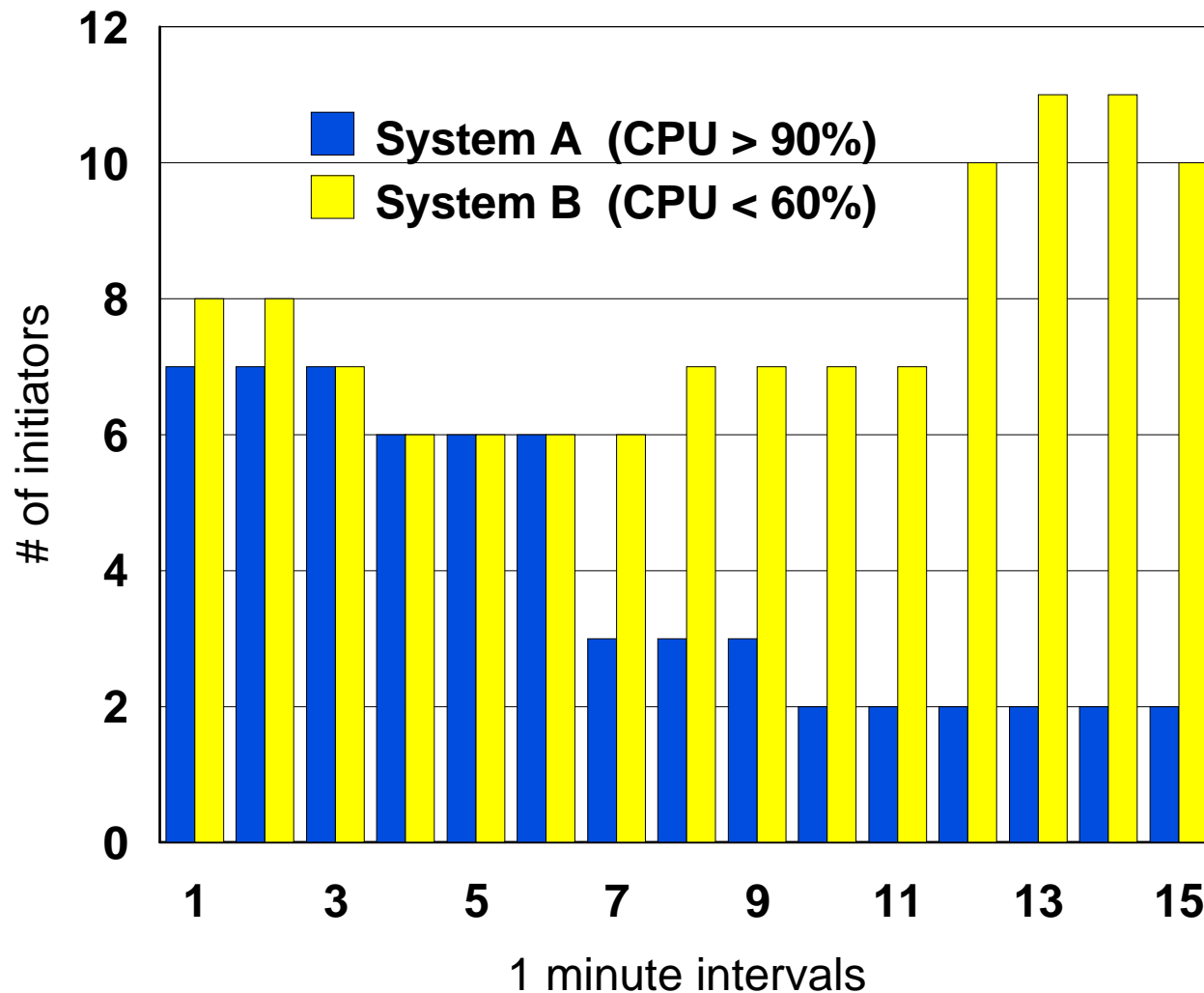
- Placement of initiators
 - Awareness of system affinities
 - Awareness of sysplex workload distribution

- Exploitation/migration on a job class basis

WLM Batch Management -- initiator placement

- Peer-peer distributed decision-making among goal mode systems
- Each system has a view of remote systems' CPU capacity, shortages, and initiator queue data
- When policy adjustment wants to add initiators, resources can be found to support them, and adding them helps measurably, try to do what a human might do:
 - Use idle capacity if available
 - If nothing idle, then proceed to
 - ▶ Have each system find which work would have to donate
 - ▶ Choose amongst the donors based on the policy goals

WLM Batch Management -- initiator placement example



- Data from a production sysplex at mid-day.
- Initiators are added on the system with more available capacity.

WLM Batch Management -- placement decision detail

Time
(secs)

System A

CPU >90%

System B

CPU <40%

0

Detect Q delay; defer to B

7

Detect Q delay; start 1 init

10

No action - remote system did it

17

No action - no receiver value

20

Detect Q delay; defer to B

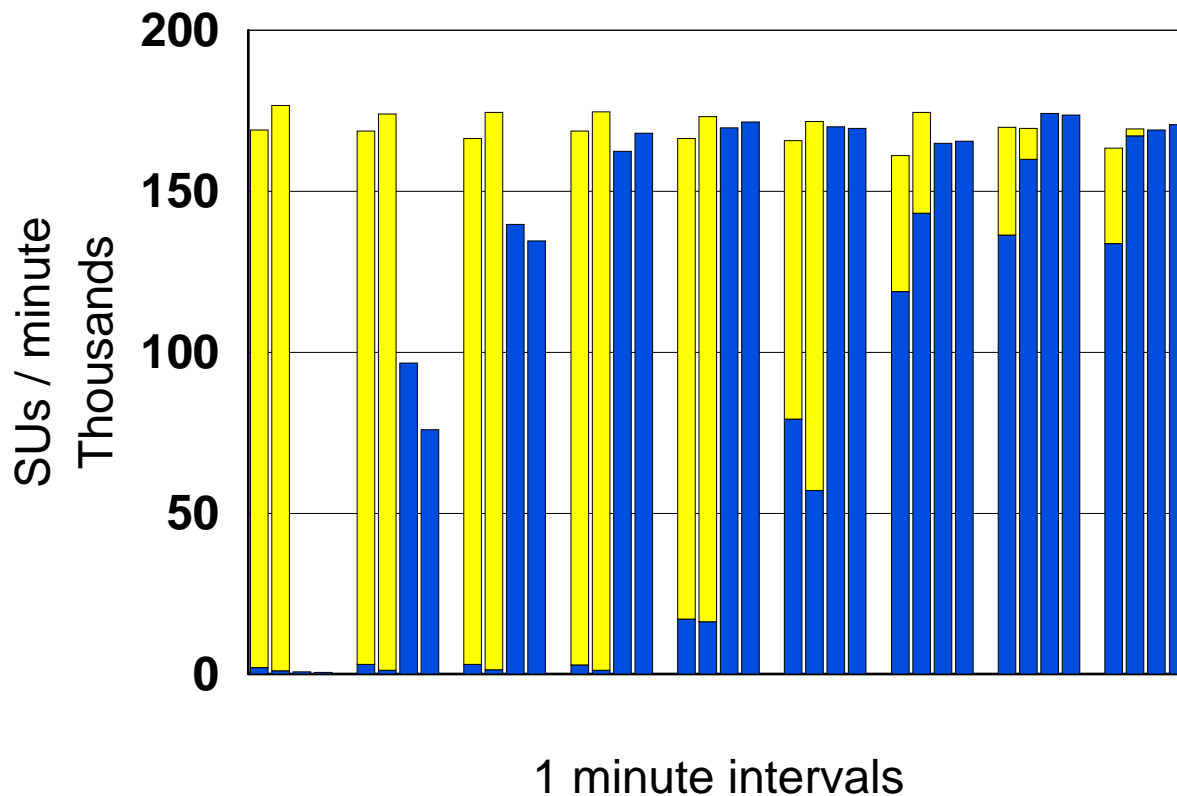
27

Detect Q delay; start 2 inits

WLM Batch Management -- 4-system example

Imp. 2 batch vs. discretionary

- discretionary
- Imp. 2 batch



- Systems A & B saturated with discretionary; C & D are idle.
- Introduce heavy Imp. 2 batch load.
- Batch uses idle capacity before displacing work.
- Displace only until batch goals are met.

Discretionary Goal Management in OS/390 V2R6

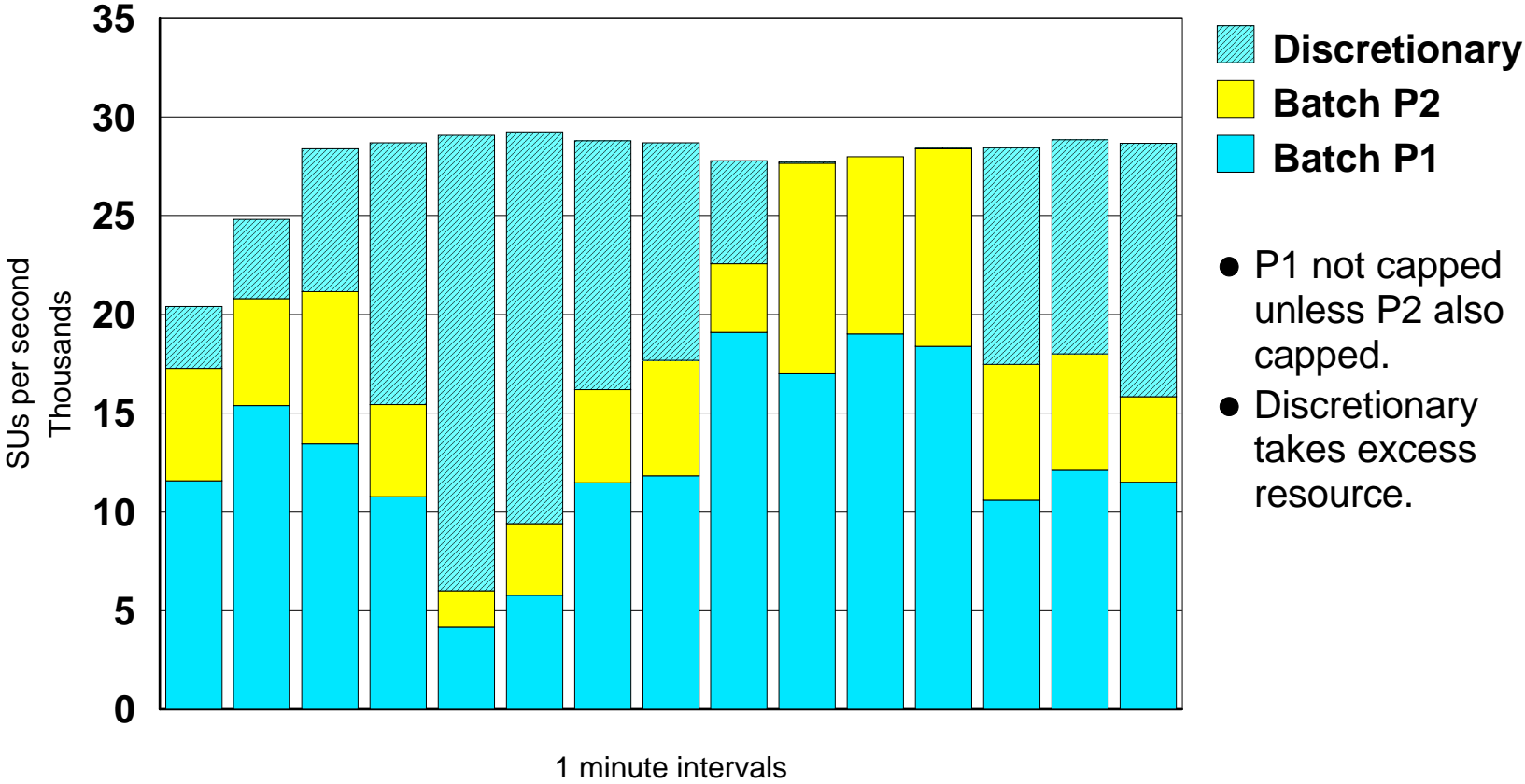
■ Pre-V2R6:

- On saturated systems, discretionary work can be shut out, even when other work is overachieving goals
- Some customers had to move discretionary work to Imp. 5
- Lose advantage of running in a mean-time-to-wait group

■ V2R6:

- Cap overachieving work to allow discretionary work to run
- Uses internal resource group maximums to manage capping
 - ▶ Capping begins when PI drops below .7
 - ▶ Capping stops when PI rises above .81
- Criteria for donors avoids impact to online work:
 - ▶ R/T goal greater than 1 minute
 - ▶ Velocity goal less than or equal to 30%
 - ▶ Later periods capped first

Discretionary Goal Management -- Example



Batch P2 capped:	-----	-----	...
PI:	.3 .3 .4 .3 .6 .6 .5 1.0 .8 .6 .5 .4 .4 .6		
Batch P1 capped:	-----	-----	...
PI:	.5 .5 .5 .8 1.0 .9 .9 .8 .9 .8 .6 .6 .7 .7		

Discretionary Goal Management

Migration Considerations for OS/390 V2R6

- Donors are now managed more closely to goals
- Revisit goals for donor service classes:
 - Have they been overachieving?
 - Do your goals reflect business requirements?
 - Does your system have unmet discretionary demand?
 - ▶ waiting for CPU
 - ▶ queued, waiting to be selected
- RMF Workload Activity report shows capping activity

Discretionary Goal Management -- Example

RMF Workload Activity Report

donor service class

REPORT BY: POLICY=O390R4H		WORKLOAD=VICOM		SERVICE CLASS=BATCH1		RESOURCE GROUP=*NONE	
TRANSACTIONS	TRANS.-TIME	HHH.MM.SS.TTT	--DASD I/O--	---SERVICE----	--SERVICE RATES--		
AVG 4.10	ACTUAL	1.36.684	SSCHRT 0.0	IOC 0	ABSRPTN 2908		
MPL 4.10	EXECUTION	17.406	RESP 0.0	CPU 7150K	TRX SERV 2908		
ENDED 110	QUEUED	1.19.277	CONN 0.0	MSO 0	TCB 1407.1		
END/SEC 0.18	R/S AFFINITY	0	DISC 0.0	SRB 69	SRB 0.0		
#SWAPS 0	INELIGIBLE	0	Q+PEND 0.0	TOT 7150K	RCT 0.0		
EXECUTD 0	CONVERSION	276	IOSQ 0.0	/SEC 11915	IIT 0.4		
	STD DEV	44.901			HST 0.0		
					APPL % 234.6		
VELOCITY MIGRATION:		I/O MGMT 9.8%	INIT MGMT 9.8%				
---RESPONSE TIME---		EX	PERF	AVG	--USING%--	----- EXECUTION DELAYS % -	
HH.MM.SS.TTT		VEL	INDX	ADRSP	CPU I/O	TOTAL	QMPL CAPP CPU
GOAL	00.03.00.000	70.0%					
ACTUALS							
SY#A		100%	9.8%	0.7	4.2	9.5	0.0 87.8 83.3 3.0 1.4
0							

goal meets criteria for donor

PI reflects capping

capping delay

Summary

WLM/SRM Internal Actions

- ▶ Impact of Importance Specification
- ▶ Use of Performance Index Reflecting Goal Attainment
- ▶ Policy Adjustment Actions and Controls

Enhancement of WLM Management Scope

- ▶ I/O Priority
- ▶ Batch Initiators
- ▶ Discretionary Goal

References

1. J. Aman, C.K. Eilert, D. Emmes, P. Yocom, & D. Dillenberger, "Adaptive Algorithms for Managing a Distributed Data Processing Workload", IBM Systems Journal, Vol. 36, No. 2, 1997. (to order, 1-800-IBM-JOUR, or contact account rep)
2. SG24-4352, System/390 MVS/ESA Version 5 Workload Manager Performance Studies