AIX Virtual User Group

http://tinyurl.com/AIXVirtualUserGroup

Active System Optimizer (ASO)

Automated AIX7 & POWER7 Tuning

- 28th June 2012

Nigel Griffiths
IBM Power Systems
Advanced Technology Support, Europe

Presentation Version 8

Abstract

- Nigel looks at the new Active System Optimizer (ASO) feature where we can use this “expert system” to autonomically & dynamically tune AIX 7 on POWER7.
- This is like having a Level 3 AIX Support performance guru tuning your system all day!
- This session includes a live demo of switching on ASO, monitoring and logging. Turns out the demo is hard to get ASO to do something interesting on demand.

- Thanks to Steve Nasypany ATS, USA and the AIX developers for ASO internals information used in this presentation

Announcement - 14th October 2011

Enhancements to IBM AIX Version 6 and AIX Version 7 offer improved performance, scalability, availability, security, and manageability

- "Active System Optimizer, a new subsystem designed to autonomically improve the performance of workloads. Performance improvements may vary depending on configuration and workload. Measurements should be taken before running the subsystem in a production environment. Active System Optimizer support is available only on POWER7® systems."
- Other performance tweaks:
  - TCP faster loopback
  - Faster rootvg WPAR Mobility
  - JFS2 dynamic changes, tuning and unmount avoidance
  - JFS2 50% reduced meta data size (AIX7 TL1 only)
- + Availability, Security, Manageability, + others

http://www-01.ibm.com/common/ssi/rep_ca/1/897/ENUS211-371/ENUS211-371.PDF

ASO Pre-Requisites

- Only AIX7.1 TL01+ on POWER7 or later
- Installed by default with AIX
  - Don't forget the mandatory Service Packs
- Warning: Any older AIX release or hardware!
  - **NOT supported**
  - May start but will do nothing

Pre-requisites Check

```
# oslevel -s
  7100-01-02-1150
  → AIX 7, TL01, Service pack 2, week 50 year 2011

# lslpp -L | grep -i optimi
 bos.aso                    7.1.1.2    C     F    Active System Optimizer

# lsconf | grep ^Processor
  Processor Type: PowerPC_POWER7
  Processor Implementation Mode: POWER 7
  Processor Version: PV_7_Compat
  Processor Clock Speed: 3108 MHz
```

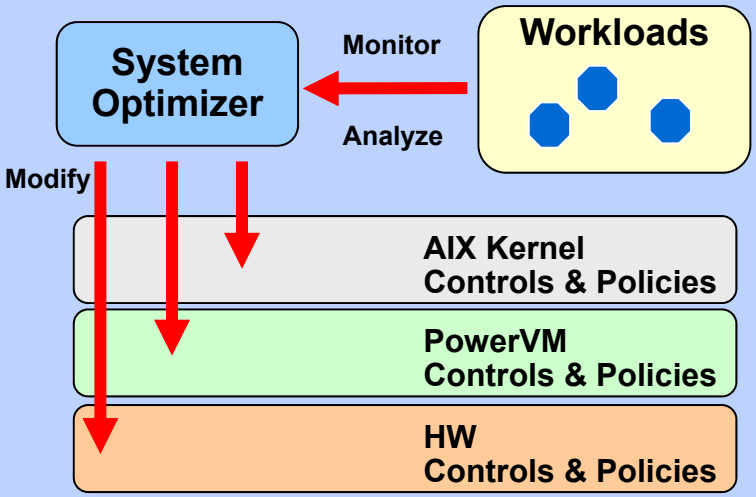
Supported configurations

- Supported:
  - AIX LPARs running in P7 compatibility mode
  - Shared Processor LPARs
    - Minimum entitlement requirement (per core and total)
  - WLM (except tiers, minimum limits)
  - WPARs, AME
- Not Supported:
  - Enhanced affinity disabled / AMS enabled
  - LPAR migration
- ASO hibernates when configuration not support

Marketing – a bit vague

- Jay "Mr AIX" Kruemcke
  - Take care - terms are vague
  - Features are being phased in but slides don't point this out
  - PowerVM + HW layers → Later
  - These may move the VM (LPAR) around the machine!!

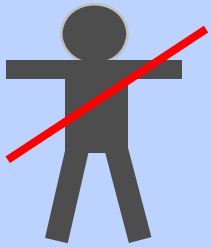
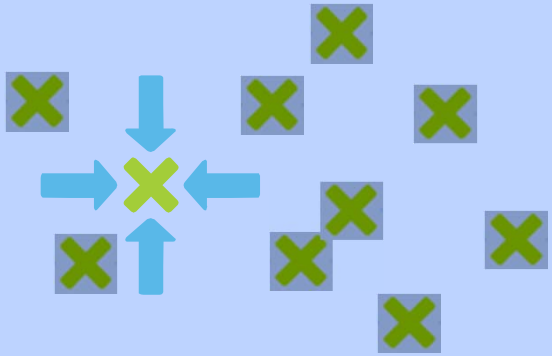
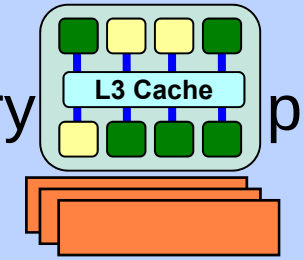
Active System Optimizer



Active System Optimizer profiles and analyses running workloads to dynamically tune system capabilities on a per workload basis

- Runtime workload monitoring and analysis
- Optimization via dynamic adjustment of policies
- Autonomic and transparent

ASO in Operation Overview

1. Once activated  requires no user involvement
2. Identifies & optimizes suitable workloads 
3. Improves cache & memory 

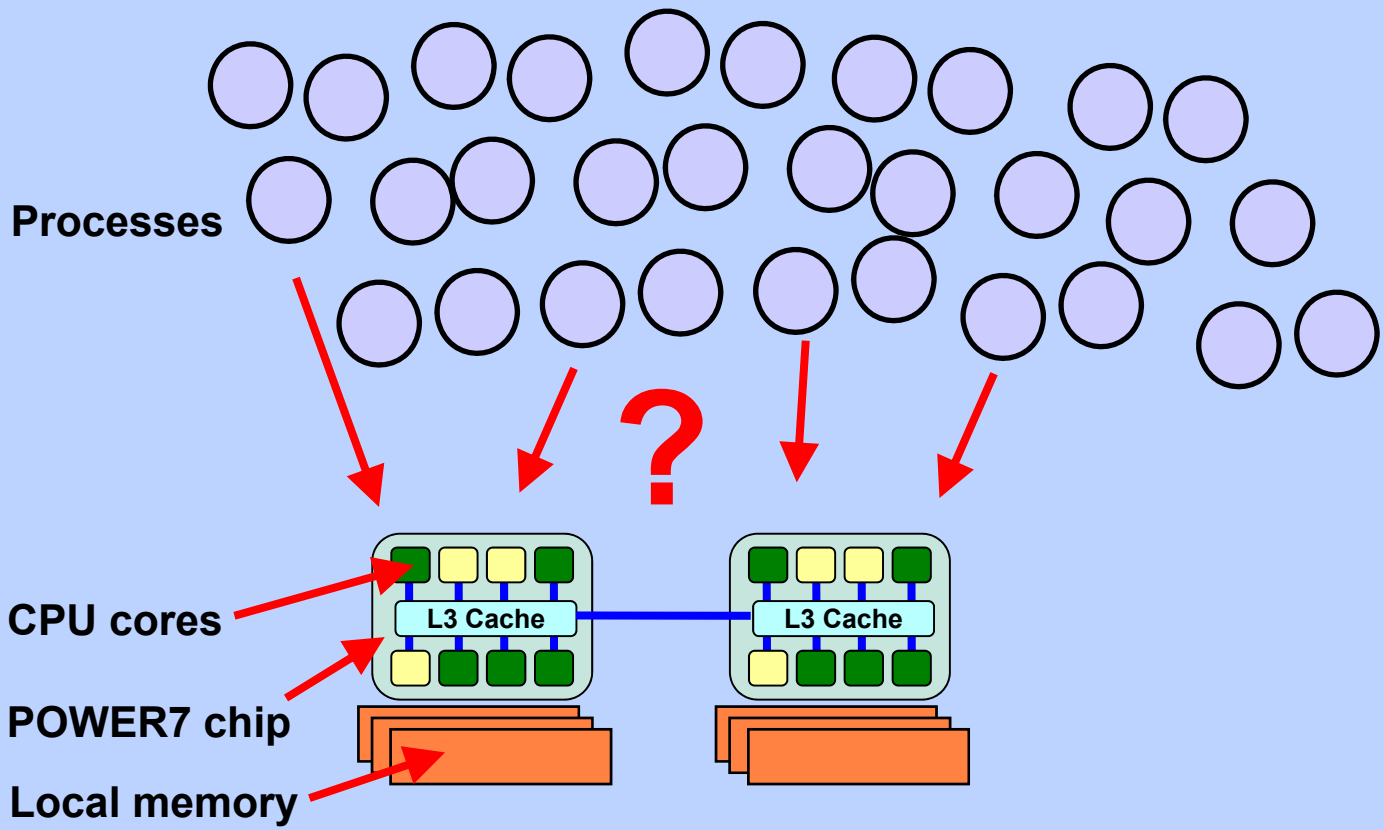
performance
4. Performs pre- & post-optimization monitoring 
5. Hibernates when not busy 

ASO in Operation Overview (in words)

- Once activated requires no user involvement
  - Autonomous and transparent
- Identifies & optimizes suitable workloads, using
  - AIX kernel data about processes/threads
  - Hardware Performance Counters from the POWER7 chips
- Improves cache & memory affinity for performance
  - Dynamically re-evaluating tuning options as the workloads change
  - Low CPU overhead, high gain in performance
- Performs pre- & post-optimization monitoring
  - Only optimizes workloads when relatively stable (minutes)
  - Adapts to changes in behaviour and workload
- Active tuning hibernates
  - If no gains achieved or unsupported environment
  - Wakes up when instrumentation indicates tuning is possible

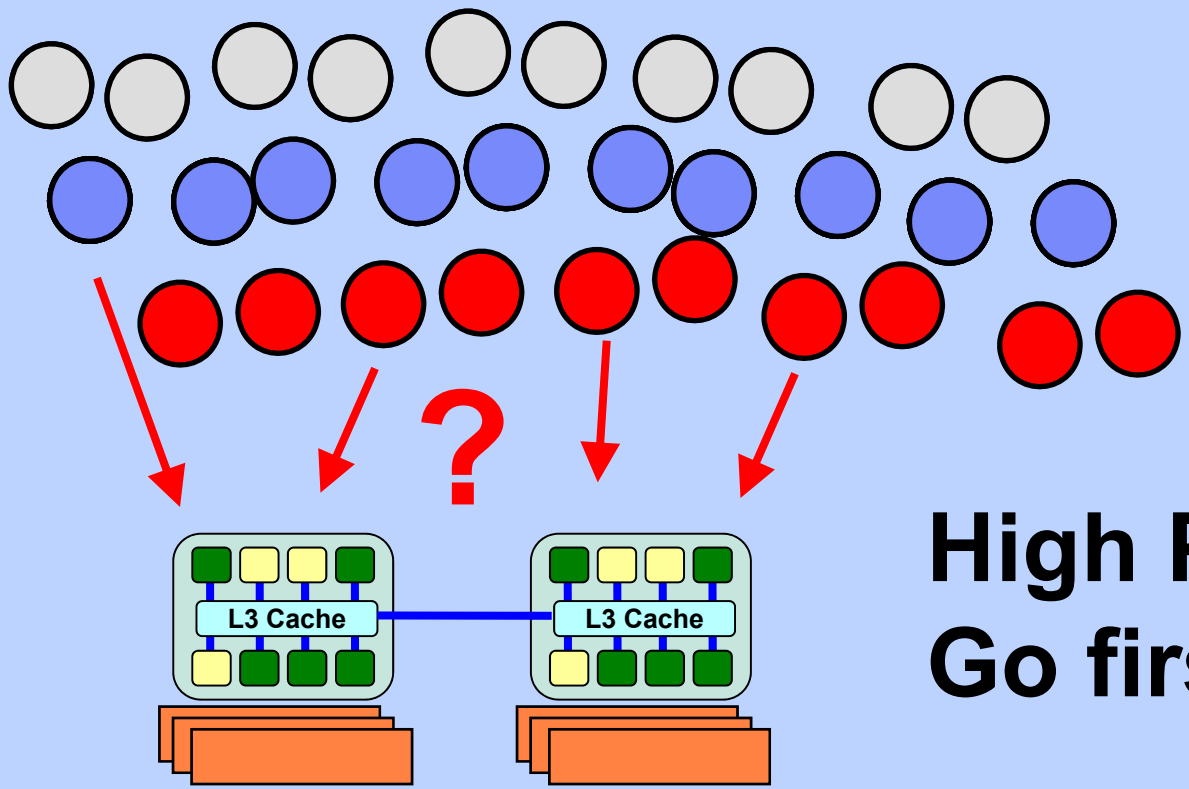
How does AIX schedule & place processes?

- AIX kernel process dispatcher = short term
  - Needs to make high speed decisions (micro seconds)

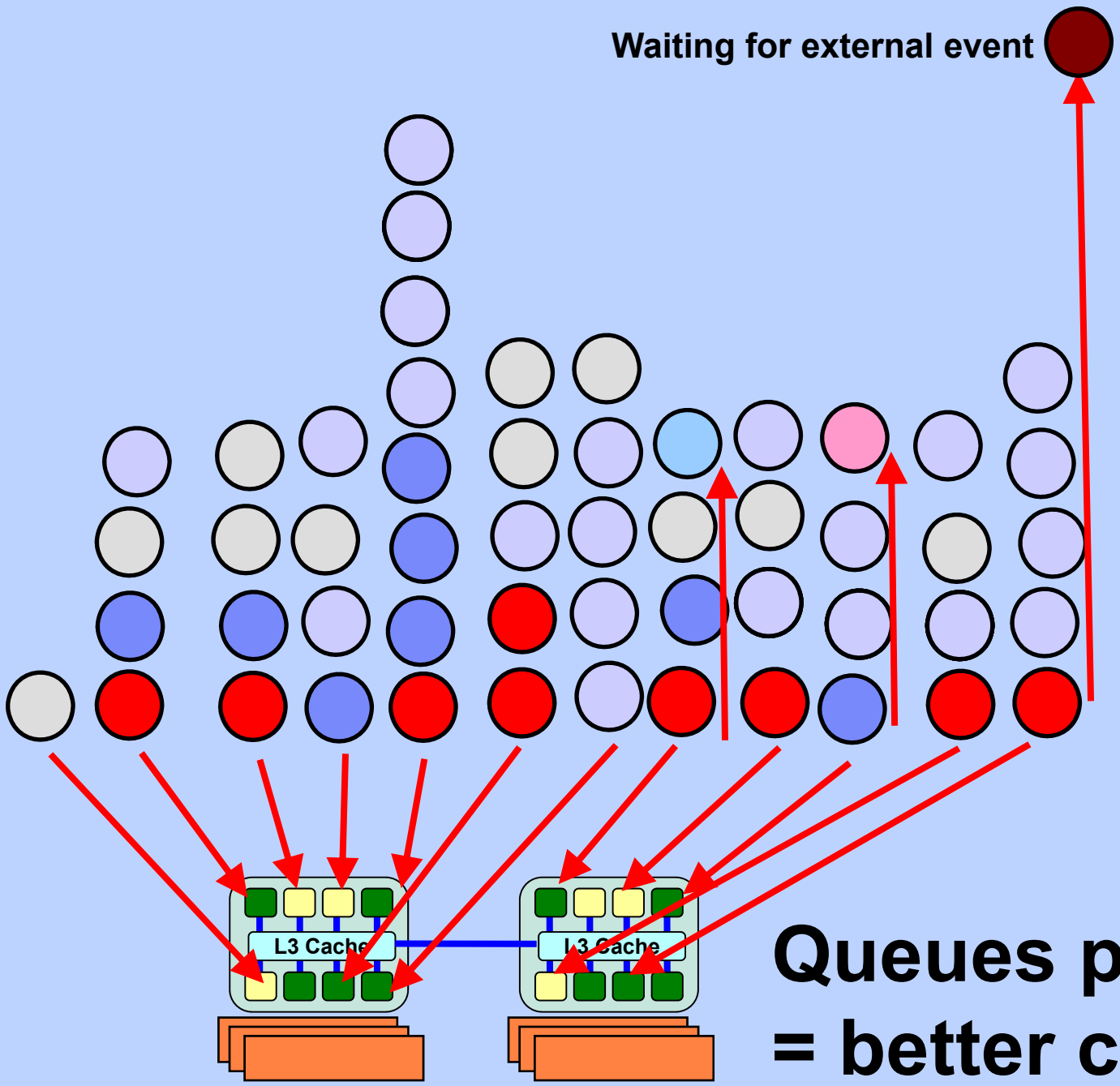


How does AIX schedule & place processes?

- AIX kernel process dispatcher = short term
  - Needs to make high speed decisions (micro seconds)

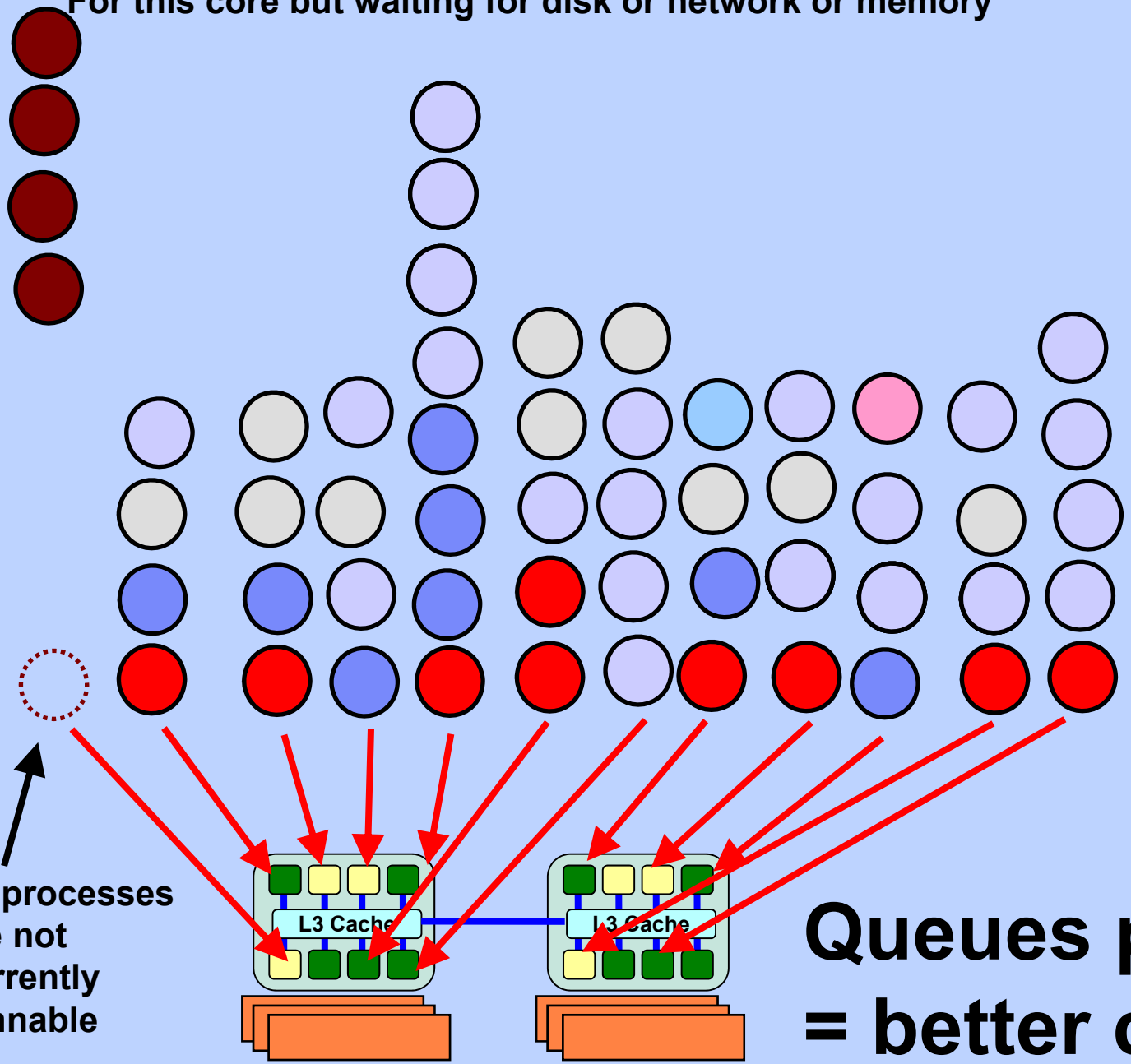


High Priority
Go first



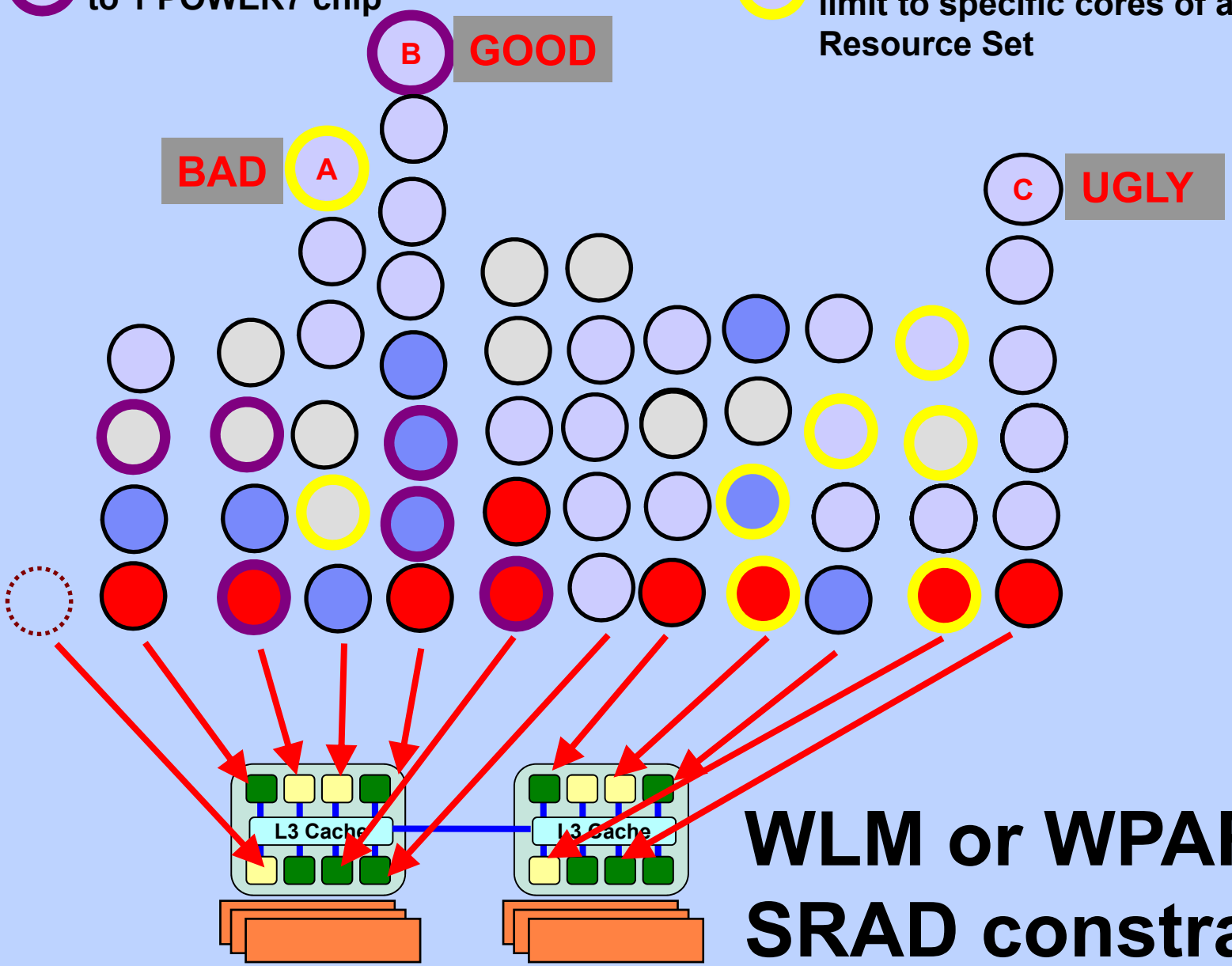
Queues per Core = better caching

For this core but waiting for disk or network or memory



Its processes are not currently runnable

Queues per Core = better caching



WLM or WPAR & SRAD constraints

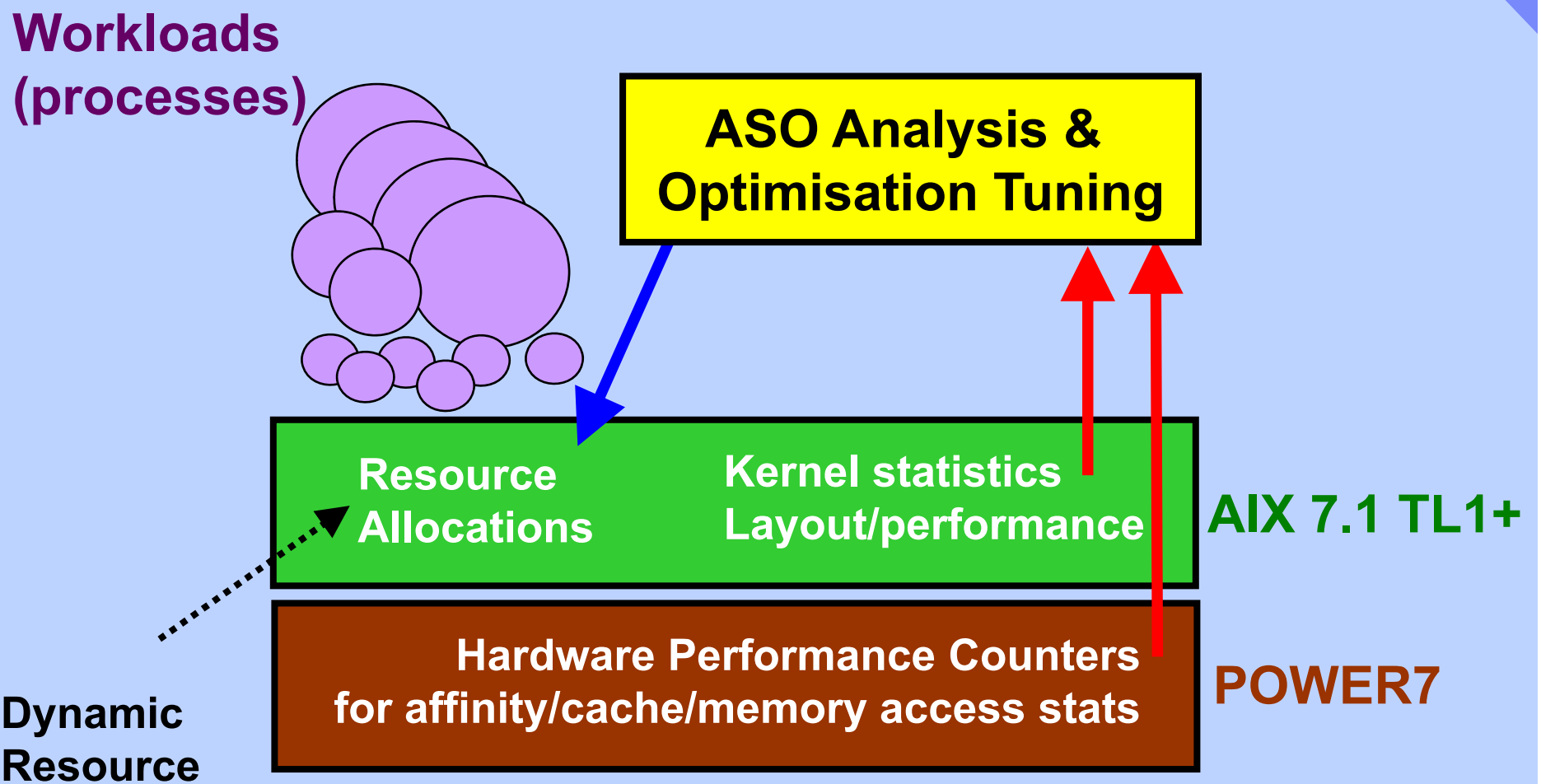
How AIX schedule & place processes?

- AIX kernel process dispatcher = short term
  - Needs to make high speed decisions (micro seconds)
  - Follows simple priority rules & queues
  - Has limited data for large-scale placement decisions
  - Potentially high cost of poor placement decision
  - Conservative by design

How ASO gets it's Tuning "Actioned" ?

- AIX kernel process dispatcher = short term
  - Needs to make high speed decisions (micro seconds)
  - Follows simple priority rules & queues
  - Has limited data for large-scale placement decisions
  - Potentially high cost of poor placement decision
  - Conservative by design
- Active System Optimizer
  - Focused on longer term analysis (minutes)
  - Time + history for better placement decisions
  - Works by setting dispatcher SRAD and RSET rules

Functional Model



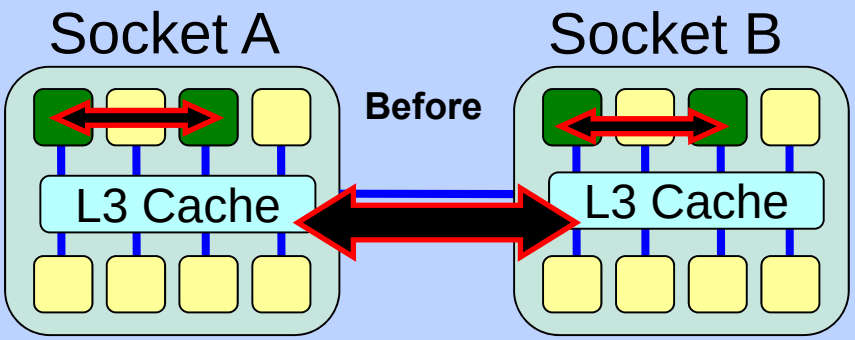
Dynamic Resource Sets (RSET) or SRAD to change the CPU a process is running on

Three types of optimisation

1. Cache Affinity
   - Reduce chip to chip cache movement
2. Aggressive Cache Affinity
   - Reduce chips involved (so less movement)
3. Memory Affinity
   - Make memory more local (less near and far access)

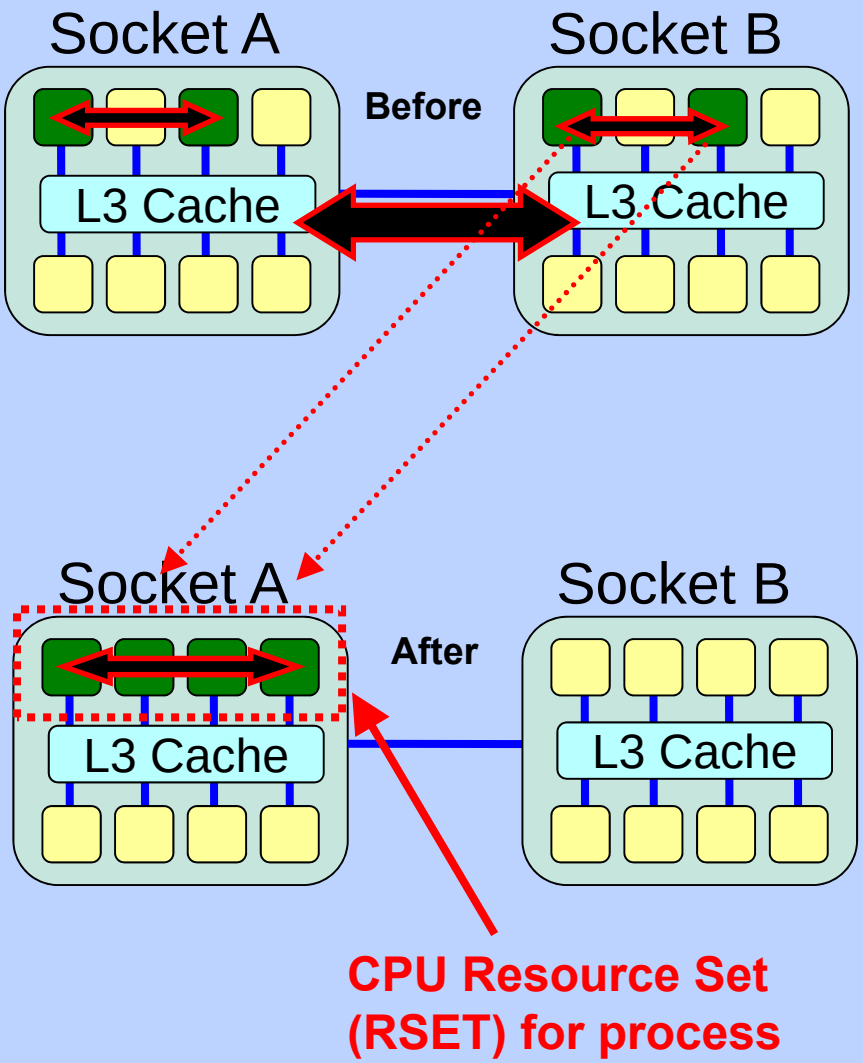
Technical Information: Optimizations

1. Cache Affinity
   - Threads of eligible workloads bound to a set of cores close together
   - Workloads monitored before and after placement
   - Load, CPU utilization, latency …
   - Conservative placement to ensure sufficient resources for workload



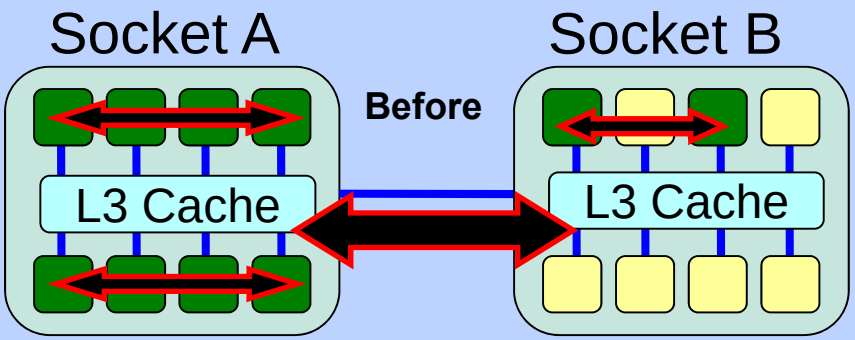
Technical Information: Optimizations

1. Cache Affinity
   - Threads of eligible workloads bound to a set of cores close together
   - Workloads monitored before and after placement
   - Load, CPU utilization, latency …
   - Conservative placement to ensure sufficient resources for workload



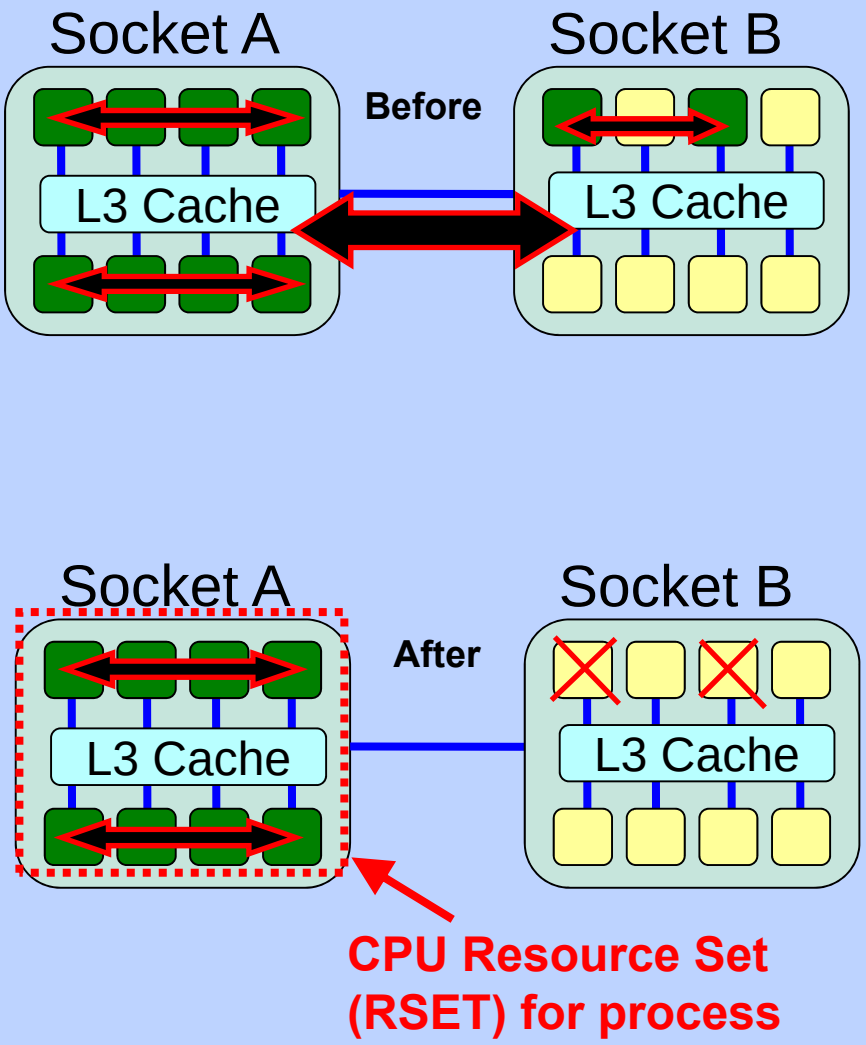
Technical Information: Optimizations

2. Aggressive Cache Affinity
   - Workloads may be compressed onto fewer cores for higher performance, profiling using PMU hardware
   - If sufficient evidence for potential improvement
   - Thorough pre- and post-optimization analysis



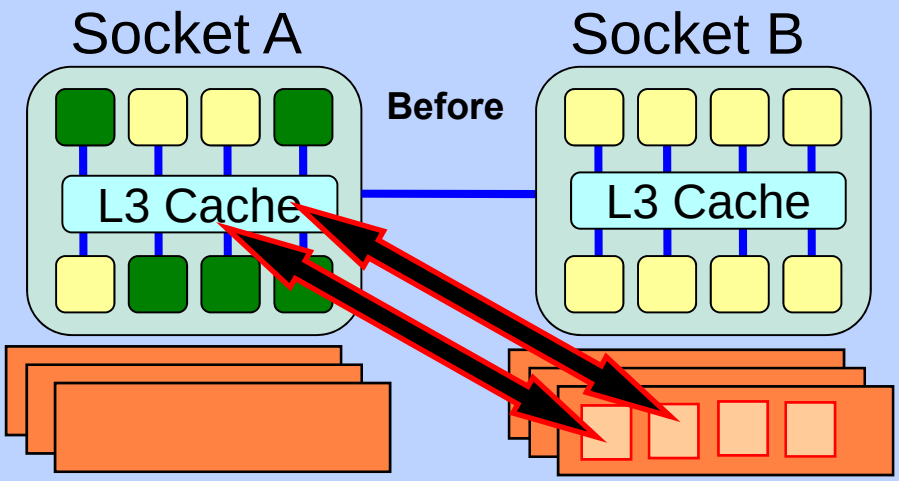
Technical Information: Optimizations

2. Aggressive Cache Affinity
   - Workloads may be compressed onto fewer cores for higher performance, profiling using PMU hardware
   - If sufficient evidence for potential improvement
   - Thorough pre- and post-optimization analysis



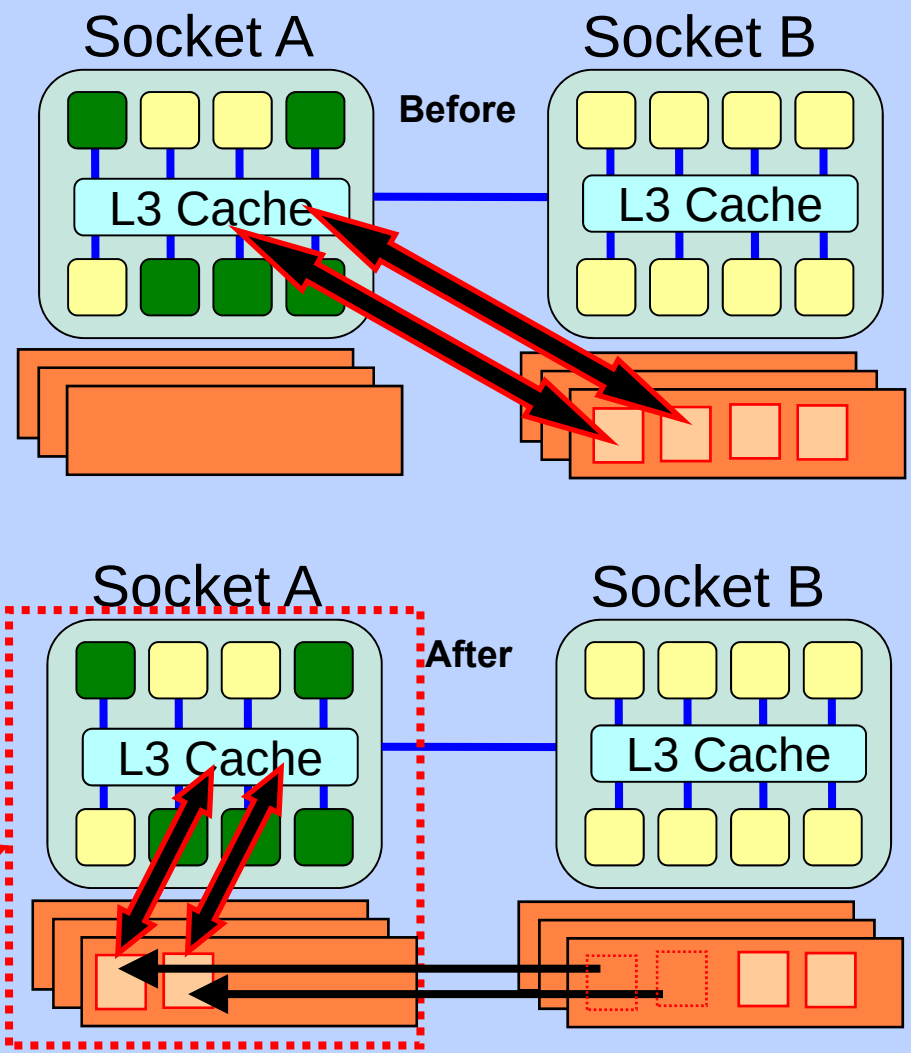
Technical Information: Optimizations

3. Memory Affinity
   - May migrate pages for workloads with high level of remote traffic
   - Incremental migration of frequently accessed pages



Technical Information: Optimizations

3. Memory Affinity
   - May migrate pages for workloads with high level of remote traffic
   - Incremental migration of frequently accessed pages



Technical Information: Eligible Workloads

- Multi-threaded workloads with periods of stability
  - CPU Utilization, Load and Latency must be stable for a period of time
- Minimum utilization = machine must be busy
  - Higher for aggressive cache optimization
- Minimum lifetime
  - 10 seconds (5 minutes for memory affinity)
- Not manually tuned
  - If too much of the system load is manually tuned, ASO hibernates
- Not explicitly marked as unoptimizable

ASO - Five things you need to know

1. ASO runs as an SRC kernel service: lssrc, startsrc
   - Must be **active**
2. Active System Optimizer Options command: asoo
   - Must be **active**
3. Other asoo Tuning options
4. Logging
   - To two simple text files
5. Fine Control of aso with Shell Variables

- Don’t confuse aso daemon with asoo tuning cmd

1) and 2) ASO Start service and Activate

1) Start the service via Systems Resource Controller :

```
# lssrc -s aso
Subsystem        Group        PID        Status
 aso                                     inoperative
# startsrc –s aso
# lssrc -s aso
Subsystem        Group        PID        Status
 aso                           1835474   active
```

… you may eventually # stopsrc –s aso

2) Then Activate (-o option and -p = permanent):

asoo -p -o aso_active=1

Active System Optimiser is now working

1) and 2) ASO Start service and Activate

```
topas_nmon                    Host=purple7          Refresh=2 secs    16:15.15
 Top-Processes-(262)    Mode=4   [1=Basic 2=CPU 3=Perf 4=Size 5=I/O 6=Cmds]
  PID      %CPU  Size   Res   Res   Res  Char RAM      Paging        Command
           Used    KB   Set  Text  Data   I/O  Use io other repage
......
 4981098    0.0  3712  3912   572  3340     0   0%    0    0    0 aso
```

```
topas_nmon          Host=purple7          Refresh=2 secs    08:58.53
 Top-Processes-(125)    Mode=1   [1=Basic 2=CPU 3=Perf 4=Size 5=I/O 6=Cmds]
  PID     Parent User        Proc Nice Pri Status proc-Flag Thrds Files  Command
             PID Id          Group    ority                 Foreground=F
4981098 2752544 root         none    41  32 Running 0x00240103  14  336   aso
```

6 - 14 threads

From my experience on 16 CPU VM
– Not seen aso use <u>any</u> CPU time
  = Less than 0.1% of one CPU

Much larger virtual machines might see some

3) asoo to configure ASO – other optons

- Standard AIX “o” tuning tool like vmo, schedo, no …
- Displays current settings (non-restricted tunables):

  # asoo –a
  aso_active = 0
- Set a value to a tunable: **asoo -o aso_active=1**
- Permanently set: **asoo -p -o aso_active=1**
- Displays help for a tunable:

  # asoo -h aso_active
  Help for tunable aso_active:
  Purpose: Disables ASO.
  Values: Default: 0 Range: 0, 1 Type: Dynamic Unit: boolean
  Tuning: 0 indicates that the ASO is disabled. 1 indicates enabled.

- Reset to default all tunables: **asoo -D**

For more info: **man asoo** - - or - - the online manual pages

http://publib.boulder.ibm.com/infocenter/aix/v7r1/index.jsp?topic=%2Fcom.ibm.aix.cmds%2Fdoc%2Faixcmds1%2Fasoo.htm

3) asoo List option details

```
# asoo -L
NAME                        CUR    DEF    BOOT   MIN    MAX    UNIT           TYPE
     DEPENDENCIES
-------------------------------------------------------------------------------
aso_active                  0      0      0      0      1      boolean          D
-------------------------------------------------------------------------------

n/a means parameter not supported by the current platform or kernel

Parameter types:
    S = Static: cannot be changed
    D = Dynamic: can be freely changed
    B = Bosboot: can only be changed using bosboot and reboot
    R = Reboot: can only be changed during reboot
    C = Connect: changes are only effective for future socket connections
    M = Mount: changes are only effective for future mountings
    I = Incremental: can only be incremented
    d = deprecated: deprecated and cannot be changed

Value conventions:
    K = Kilo: 2^10       G = Giga: 2^30       P = Peta: 2^50
    M = Mega: 2^20       T = Tera: 2^40       E = Exa: 2^60
#
```

Restricted option – Only use if told to by support

```
# asoo -FL
NAME                        CUR    DEF    BOOT   MIN    MAX    UNIT           TYPE
     DEPENDENCIES
-------------------------------------------------------------------------------
aso_active                  0      0      0      0      1      boolean          D
-------------------------------------------------------------------------------
##Restricted tunables
-------------------------------------------------------------------------------
aggressive_cache_affinity 1      1      1      0      1      boolean          D
-------------------------------------------------------------------------------
aggressive_cache_opt_utilisation
                            1000   1000   1000   1      2000   1/1000th cores   D
-------------------------------------------------------------------------------
allow_fp_placement          1      1      1      0      1      boolean          D
-------------------------------------------------------------------------------
allow_sub_srad_placement    1      1      1      0      1      boolean          D
-------------------------------------------------------------------------------
max_placement_rate_per_srad
                            25     25     25     0      100    percent          D
-------------------------------------------------------------------------------
memory_affinity             1      1      1      0      1      boolean          D
-------------------------------------------------------------------------------
message_facility            12     12     12     0      23     numeric          D
-------------------------------------------------------------------------------
min_utilisation_dedicated   100    100    100    1      2000   1/1000th cores   D
-------------------------------------------------------------------------------
min_utilisation_shared      100    100    100    1      2000   1/1000th cores   D
-------------------------------------------------------------------------------
percent_system_to_optimise
                            80     80     80     0      100    percent          D
-------------------------------------------------------------------------------
```

Do NOT touch

4) Active Systems Optimizer Logging

ASO logging found in **/var/log/aso/***

- Format is not documented but readable ASCII text
- aso.log
  - On/Off status
  - ASO hibernate reasons like VM not busy!
  - Or tuning made things worse, manual tuning found etc.
- aso_process.log
  - Details of actions
  - Processes modified
- Hint:
  - You need to find the interesting processes that you think need tuning
  - Search for the process name to find the PID → in the [ ]
  - Then search for the PID for all the messages

5) aso – Fine Control via Shell Variables

- Warning: aso manual page
  - Says starting aso outside SRC OK but really only for debugging aso
  - But also includes Shell Variables to fine control = Good
- For more information: man aso

- Not normally needed

A bit Catch 22

- Set these before starting important processes
  - Master switch
    - ASO_ENABLED=[ALWAYS | NEVER]
  - Prioritise or stay clear of process
    - ASO_OPTIONS=ALL=[ON | OFF]
    - ASO_OPTIONS=CACHE_AFFINITY=[ON | OFF]
    - ASO_OPTIONS=MEMORY_AFFINITY=[ON | OFF]

a priority to optimise

ASO in Practice

```
# startsrc -s aso
# asoo -p -o aso_active=1
# tail -f /var/log/aso/aso_process.log
```

Workloads running ….

expect ASO to monitor workloads for a few minutes

Note:
Log format is not documented but fairly readable

Some guess work in the following example logs

… your mileages will vary as every workload is different

ASO in Practice

My VM (LPAR) cleverly badly laid out on a 2 Drawer Power 770

```
# lssrad -av
REF1   SRAD        MEM      CPU
0
          0    6958.40      0-3 8-11 16-19 28-31
          3     498.00
1
          1    5894.56      4-7 12-15 20-23
          2    1992.00      24-27
```

Below are logging extracts

– Please don’t embarrass me with ANY questions !!!

ASO in Practice on VM called purple7

Jan 17 11:39:21 purple7 aso:notice aso[4981098]: ASO enabled by tunable
Jan 17 11:39:26 purple7 aso:notice aso[4981098]: [WLM] Is now active.
Jan 17 11:39:46 purple7 aso:notice aso[4981098]: [HIB] SPLPAR local dispatch ratio is below threshold (37%).
Jan 17 11:39:46 purple7 aso:notice aso[4981098]: [HIB] At least 50% of VCPU dispatches must be local to run ASO

Jan 17 11:41:41 purple7 aso:notice aso[4981098]: [HIB] Resuming from hibernation.

Removing the Date Time VM name process for readability

ASO enabled by tunable ← ASO started with asoo
[WLM] Is now active.
[HIB] SPLPAR local dispatch ratio is below threshold (37%).
[HIB] At least 50% of VCPU dispatches must be local to run ASO
← No work so ASO hibernated
[HIB] Resuming from hibernation. ← Work started

Hibernation event

ASO in Attempted Optimisation

[perf_info] system utilisation 0.00; total process load 0.00

[SC][5374024] Considering for optimisation (cmd='paraworms', utilisation=4.14, pref=0; attaching StabilityMonitorBasic)
[EF][6226514] attaching strategy StabilityMonitorAdvanced

[HIB] SPLPAR local dispatch ratio is below threshold (12%).
[HIB] At least 50% of VCPU dispatches must be local to run ASO
Process 5374024 (paraworms): Resetting optimisation
[SC][5374024] Removing strategy StabilityMonitorBasic from job

paraworms program looks interesting so monitors it to ensure it is not a transitory peak or short lived process

But it fails on this criteria so monitoring stops & hibernate

ASO in Optimisation Negative Effect

Suggested SRAD change

```
[SC][6226514] Considering for optimisation (cmd='paraworms', utilisation=1.62, pref=0;
   attaching ExperimenterStrategy)
[EF][6226514] attaching strategy ExperimenterStrategy
[EXP] Allowing domain System
[PRED][6226514] SRAD (2): -Cross: 0.00 +Compr: 6.40 Gain: -6.39  -- SCORE: 0.75
[PRED][6226514] Book (2): -Cross: 0.00 +Compr: 6.40 Gain: -6.40  -- SCORE: 0.75
[PRED][6226514] Recommending max domain None of minimum size 68
[EXP][6226514] Predictor recommends trying None (68)
[EXP][6226514]: giving up experimenting because only 1 domains allowed.
[EXP][6226514] Detaching without recommendation.
[PRED][6226514] SRAD (2): -Cross: 0.00 +Compr: 6.46 Gain: -6.46  -- SCORE: 0.74
[PRED][6226514] Book (2): -Cross: 0.00 +Compr: 6.46 Gain: -6.46  -- SCORE: 0.74
[PRED][6226514] Recommending max domain None of minimum size 68
[EF][6226514] detaching strategy ExperimenterStrategy
[SC][6226514] Removing strategy ExperimenterStrategy from job
[EF][6226514] detaching strategy PredictorStrategy
[SC][6226514] Removing strategy PredictorStrategy from job
[perf_info] system utilisation 1.58; total process load 2.99
[EF][6226514] clearing timeout for strategy StabilityMonitorBasic
[EF][6226514] clearing timeout for strategy StabilityMonitorAdvanced
```

Negative gain
so thinking again

ASO in Optimisation Using SRADs

Three multi-thread apps running

```
[SC][7012560] Considering for optimisation (cmd='paraworms', utilisation=2.23, attaching StabilityMonitorBasic)
[SC][1835312] Considering for optimisation (cmd='paraworms', utilisation=1.18, attaching StabilityMonitorBasic)
[SC][6226514] Considering for optimisation (cmd='paraworms', utilisation=1.17, attaching StabilityMonitorBasic)
[perf_info] system utilisation 4.71; total process load 9.96
attached(   7012560): cores=4, firstCpu=  0, srads={0}
[WP][7012560] Placing non-FP (norm load 3.20) on 4.00 node
attached(   1835312): cores=3, firstCpu=  4, srads={1}
[WP][1835312] Placing non-FP (norm load 2.40) on 3.00 node
[EF][sys_action][7012560] Attaching (load 3.20) to domain SRAD (cores=4,firstCpu=0)
[EF][sys_action][1835312] Attaching (load 2.40) to domain SRAD (cores=3,firstCpu=4)
[perf_info] system utilisation 5.24; total process load 9.96
[perf_info] system utilisation 4.91; total process load 9.93
[SC][7012560] Considering for optimisation (cmd='paraworms', utilisation=1.85, attaching StabilityMonitorAdvanced)
[EF][7012560] attaching strategy StabilityMonitorAdvanced
[SC][6226514] Considering for optimisation (cmd='paraworms', utilisation=1.39, attaching PredictorStrategy)
[EF][6226514] attaching strategy PredictorStrategy
[SC][1835312] Considering for optimisation (cmd='paraworms', utilisation=1.37, attaching StabilityMonitorAdvanced)
[EF][1835312] attaching strategy StabilityMonitorAdvanced
[perf_info] system utilisation 4.61; total process load 9.96
[EXP] Allowing domain System
[PRED][6226514] SRAD (4): -Cross: 0.00 +Compr: 0.00 Gain: 0.00  -- SCORE: 1.00
[PRED][6226514] Book (4): -Cross: 0.00 +Compr: 0.00 Gain: 0.00  -- SCORE: 1.00
[PRED][6226514] Recommending max domain SRAD of minimum size 4
[EXP][6226514] Predictor recommends trying SRAD (4)
[EXP] Allowing domain Book (4)
[EXP] Allowing domain SRAD (4)
attached(   1835312): [free]
[EF][sys_action][1835312] Detached from rset
[HIB] SPLPAR local dispatch ratio is below threshold (12%).
[HIB] At least 50% of VCPU dispatches must be local to run ASO
Process 6226514 (paraworms): Resetting optimisation
```

2 assigned different SRADs
i.e. cores on different chips

Unset 1 app as Gain = 0 = no improvement

ASO in Practice

My VM (LPAR) cleverly badly laid out on a 2 Drawer Power 770

```
# lssrad -av
REF1   SRAD        MEM      CPU
0
          0    6958.40      0-3 8-11 16-19 28-31
          3     498.00
1
          1    5894.56      4-7 12-15 20-23
          2    1992.00      24-27
```

Below are logging extracts

– Please don’t embarrass me with ANY questions !!!

Troubleshooting

Note: ASO places entire processes
- All threads of a process are currently treated the same
- Does not group threads within processes

- If ASO does not yield expected performance boost:
  - Check log file, search for process names then search for the PID
  - Copy suspect log files, prepare a perfPMR capture with & without ASO
  - Define your expected result to report in a PMR

- Manual tuning comparisons
  - This assumes Guru level Affinity & WLM & RSET skills are available!
  - Use SRAD/RSET attachments/CPU bindings (e.g. execrset / attachrset)
  - Memory affinity environment variable (MEMORY_AFFINITY=MCM)
  - Manual tuning may require constant supervision & continuous adjustments!

Performance Benefits

- Out of box performance boost for many workloads
  - *Multi-threaded, memory / cache intensive, poor scaling*

- Example workloads
  - *SpecJBB – multi-threaded JVM benchmark*
    - *From 16 cores (2 sockets) up to 72 cores (9 sockets)*
  - *Daytrader – Websphere (java) + DB2*
    - *16 / 32 cores (2 / 4 sockets)*
  - *Websphere Message Broker (WMB)*
    - *16 cores*
  - *COPR – large DB2 benchmark*
    - *64 cores (8 sockets)*

Benchmark	SpecJBB	Daytrader	WMB	COPR
ASO	Banned			
Hand Tuned				

Active System Optimizer Summary

1. AIX 7.1 TL01+ on POWER7 or later = "Set & forget"

2. Advanced Autonomic Affinity Tuning
   - Low CPU impact with zero negative effects
   - High performance boost

3. Particularly good for
   - Complex, multi-threaded, long running processes
   - Large CPU + RAM LPARs on larger machines

First phase Optimiser … with more to come