Performance Matters: Capacity measurement, limiting, and automation opportunities

CMG Canada: Spring 2014

John Baker
jbaker@mvssol.com
www.mvssol.com
Agenda

- The Four Hour Rolling Average (4HRA) and implications
- Capacity limits and implications
- When, where and how can automation be applied?
- Questions and discussion
Since 1988
- 200 z/OS customer installations
- IBM Partner in Development

The Leader in z/OS Batch Innovation
- ThruPut Manager, Version 7.1
- SE, AE, AE+
- Batch automation across the enterprise
22 years Mainframe experience focusing on performance
- Hardware and software tuning, WLM, GDPS...
- Enterprise I/O performance
- Experience on both the customer and vendor/consultant sides
The Four Hour Rolling Average
4HRA Defined

- A number representing the average utilization of the LPAR over the previous four hours
- Recorded in the SMF type 70 (SMF70LAC) for each LPAR
- Value is the rolling 4-hour average utilization of the LPAR at the end of the interval
- Unit is Millions of Service Units per hour (MSU)
- All IBM System z machines have a published MSU rating (see LSPR)
4HRA: A Large Auto-maker
4HRA: Primary purpose

- Traditional IBM Monthly License Charges (MLC) are based on the capacity of the entire machine.
- Sub-Capacity Pricing model charges based on the monthly 4HRA peak.
- Sub-Capacity Reporting Tool (SCRT) combines SMF 70 data (4HRA) with SMF 89 (what products are active) to produce a Sub-Capacity Report which is sent to IBM.
- IBM uses this report to generate MLC invoice.
A method of limiting the maximum consumption of MSU’s on an LPAR or group of LPAR’s

Set via Hardware Management Console (HMC)

Variations are:

- Initial Cap: Hard cap – no averaging
- Defined Capacity: LPAR scope. Uses 4HRA
- Group Capacity: Shared limit for group of LPAR’s on a single CEC
Capacity Limit

4HRA with Capacity Limit

- Interval Demand
- 4HRA
- Capacity Limit
Capacity Limit Implications

- **Initial Cap (Hard cap)**
  - May not be exceeded – even on an interval basis
  - Precludes soft cap options (below)

- **Defined Capacity**
  - Interval (burst) usage may exceed cap
  - If 4HRA exceeds cap, WLM will inform PR/SM to cap the LPAR until the 4HRA is at or below the cap

- **Group Capacity**
  - Same principle as Defined Capacity but there is an overall (shared) limit for every LPAR in the group
  - May be combined with Defined Capacity
Hard Capping: Inflexible

Inflexible unless limit is very modest
Soft caps allow the interval demand to exceed the limit but the LPAR(s) will be capped by PR/SM if the 4HRA exceeds the limit.
When the 4HRA is exceeded, WLM instructs PR/SM to cap the LPAR. Note the immediate effect on the interval consumption.
Automation
Managing Caps manually

Monitors and displays to be constantly aware of the changing workload demand and resource use

Potentially hours of fire fighting

Recovering; resetting back to normal; loop back to monitoring
Automation options

- Automate resource provisioning – avoid hitting wall
How to maintain the cap?
What about automating demand?

- Is there a way to reduce or delay some workloads without negative impact?
- What are the candidates? Interactive or Batch

**Batch:**
- Generally lower priority
- Does not require sub-second response
- Can be controlled via initiators
- Once executing, can be controlled via Resource Groups

- Contributes to total demand and can be manipulated
4HRA peaks tend to occur during the day. While online is typically the largest consumer, we find approximately 30% of the workload is batch.
How to automate?

- Classify Batch by priority – review WLM Policies
- Begin Do loop...
  - Monitor system metrics – ongoing using commands, monitors
    - 4HRA relative to cap level
    - Interval demand
  - Drain/halt initiators as peak approaches – using commands
  - Reset Service Classes for executing work (use RG limits) – using commands
  - Undo constraints as peaks pass – commands
- Continue to Begin loop -
- every 10 seconds...
- Or get real automation software – it’s out there!
ThruPut Manager Automation Edition

- Automates management of all batch based on installation goals and limits
- Gradually constrains low priority batch (as designated by you) as 4HRA peak approaches. Releases as peaks pass
- Result:
  - Allows lower cap level = lower $$MLC
  - Protects loved ones
Automated Demand

Caps can be lowered while protecting high priority workloads.
Thank You!

Questions/discussion