

ThruPut
Manager AE

Getting the Best
Out of
Sub-capacity Pricing

Spring 2011

© 2011 MVS Solutions Inc.

MVS
solutions inc.

Why is MVS Solutions talking about this?

Because we've become aware of the effect of batch on pricing and we're working in this area.

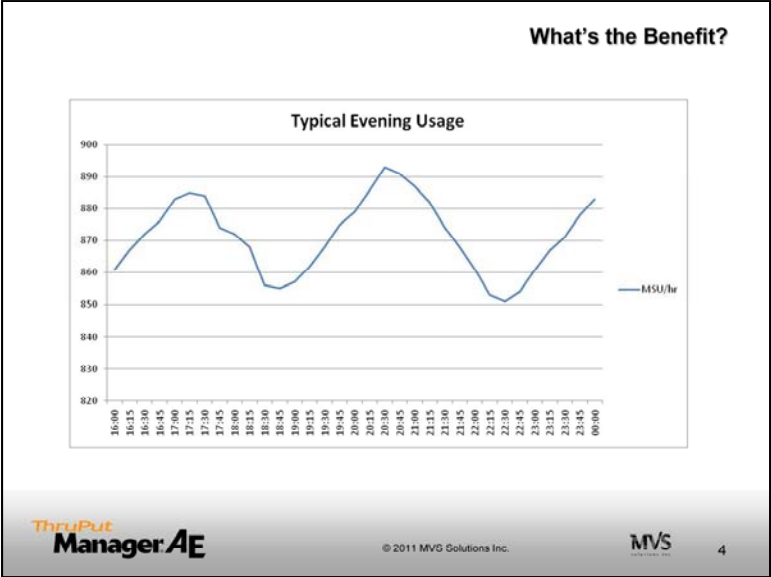
Agenda

- Software charging
- The 4-hour rolling average
- Soft capping
- Impact of capping
- Impact of batch
- Our contention

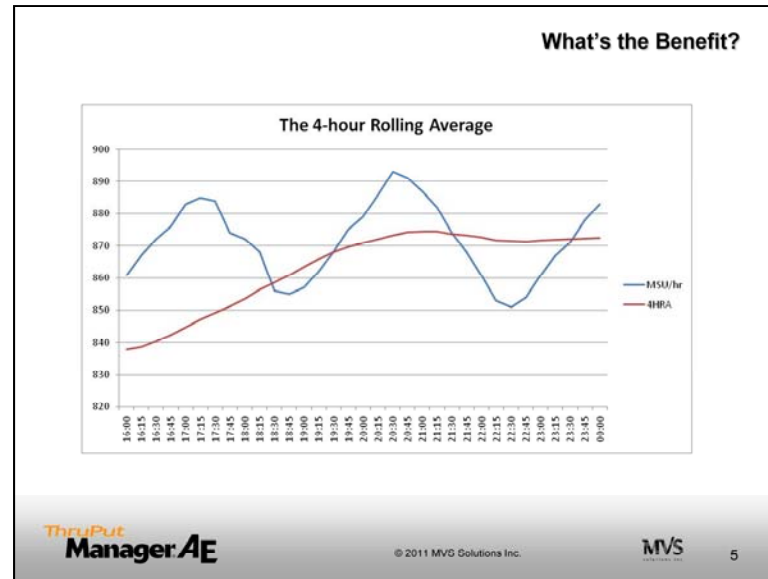
I'm not really going to talk about our product other than our contention at the end.

Software Charging

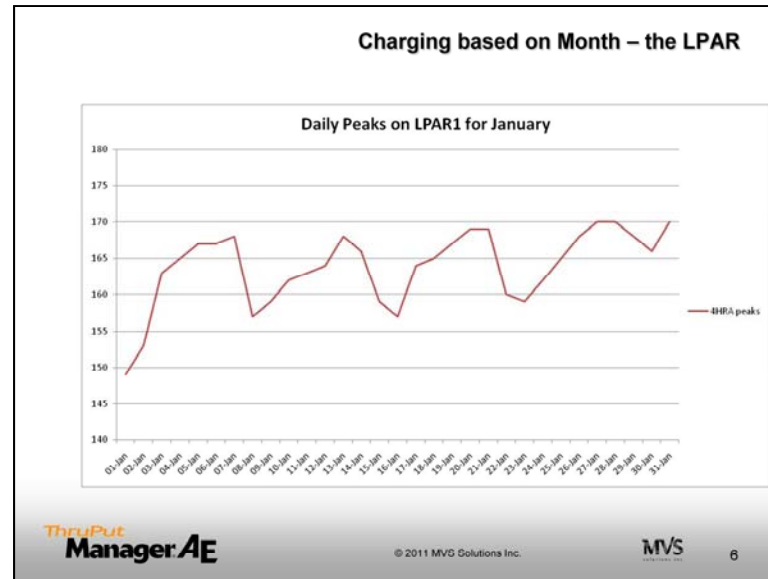
- VWLC and Sub-capacity AWLC
- Pay for software by usage of the LPAR or CEC, not capacity
- Based on highest usage of the month, as calculated by the average over a four-hour period
- Eligible software
 - Most IBM system software products – z/OS, CICS, DB2
 - Many other IBM products – Cobol, ...
 - Few other vendors



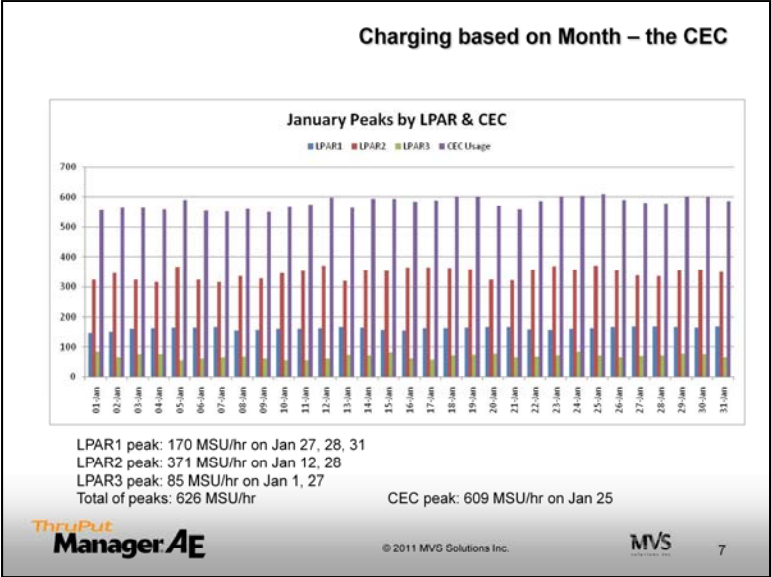
Here's a typical 8-hour period showing the MSU/hr usage pattern. As you can see there are peaks up around 890 level and valleys down to about 850 on a machine with a capacity of 900.



4HRA determines the capacity for charging purposes.
 As you can see, the 4HRA is much lower than the peaks due to the averaging.
 But charging is not based on an 8-hour period but on the month.



Here's a month's figures for an LPAR.
 Dips on weekends, peaks near month-end
 But charging typically not based on an LPAR but rather the CEC



Here's a CEC showing the load of each of its 3 LPARs and the total load.
 You can see the monthly peaks for each LPAR.
 Charging is not based on the sum of the peaks but on the 4HRA peak for the entire CEC.

Controlling the Costs
LPAR Level Soft Capping: Defined Capacity

- Defined Capacity provides an MSU/hr charging limit for the 4HRA on an LPAR running z/OS native
- Set in the HMC and can be changed dynamically
- Good for products licensed for a single z/OS LPAR on this CEC
 - May minimize costs for those products
- Not so good for general usage
 - May leave cycles on the table

If the LPAR is capped – cannot use any more – while other LPARs have a light load, there will be unused cycles you're paying for. You can't get them back.

Controlling the Costs
CEC Level Soft Capping: LPAR Group Limit

- LPAR Group Limit provides an MSU/hr charging limit for the 4HRA for a group of z/OS LPARs on the same CEC
 - Only for partitions with shared CPs
 - Not compatible with hard capping
 - May be in different Sysplexes
- Set in the HMC
- Managed by WLM
- IRD will not adjust weights while capping is effective

I call it a charging limit because even though your MSU/hr and your 4HRA may exceed this level, this is what you pay for.

IRD can be effective at adjusting weights but is effectively turned off when capping comes into effect.

CEC Level Soft Capping: LPAR Group Limit

- Good for products licensed on all z/OS LPARs
- Makes best use of the available cycles
 - Provided weights are set correctly
- LPAR Group Limit may be combined with Defined Capacity
 - The lower capping level will be used

Since charging for most products is based on the CEC, it makes most sense to put the limit at this level. But as you'll see, weights provide the basis for each LPARs apportionment when capping is in effect.

How Does Soft Capping Work with Defined Capacity?

- Controlled by weight
- But, weights are not changed by capping
 - If the weight is smaller than the Defined Capacity, a cap pattern is defined that caps at its weight then uncaps
 - Can be drastic fluctuations if the gap is large
 - If the weight is larger than the Defined Capacity, a 'phantom weight' is defined
 - Pretends there is an extra LPAR in the mix and thereby lowers the weight percentage for the current LPAR

➤ z/OS MVS Planning: Workload Management

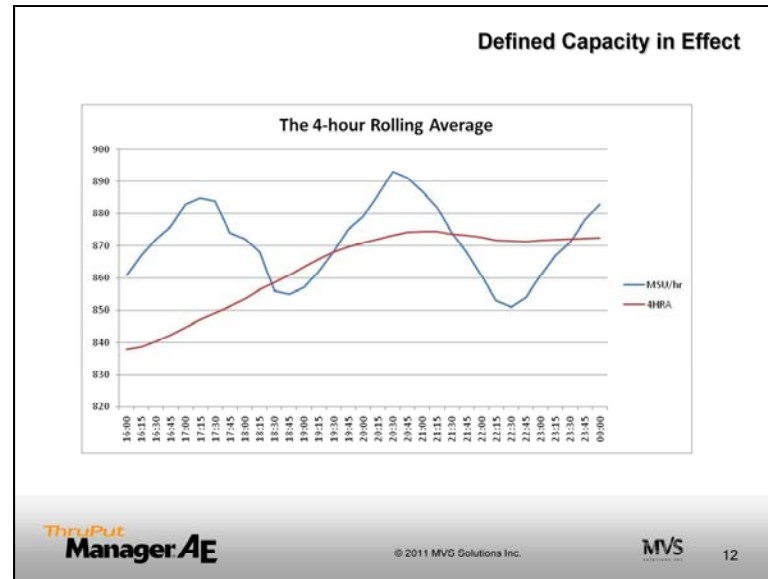
ThruPut
Manager AE

© 2011 MVS Solutions Inc.

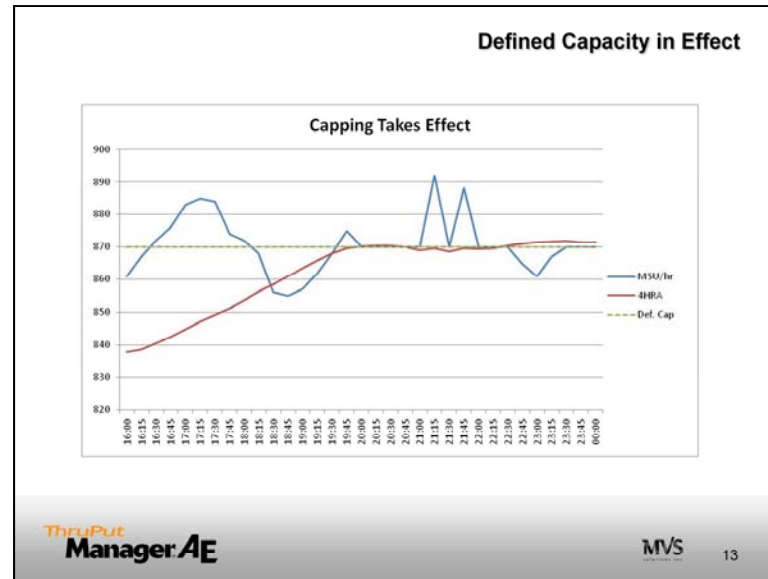
MVS

11

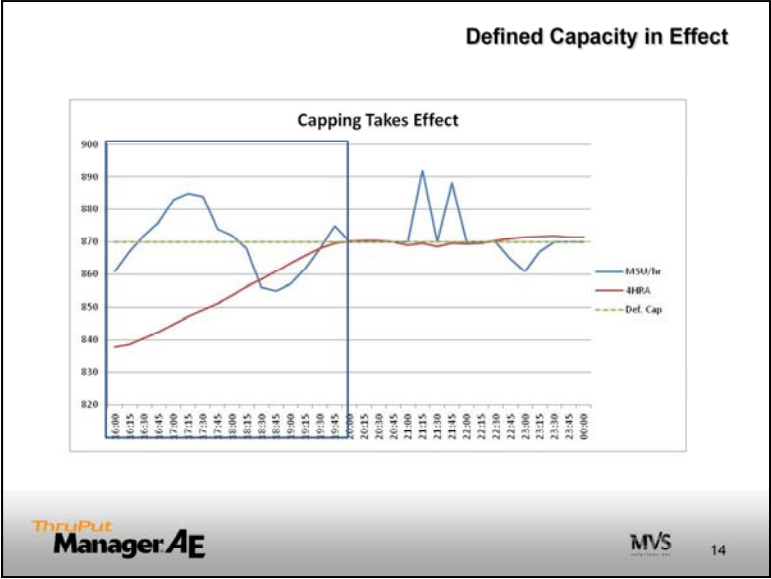
Even though weight provides the basis for doling out cycles, and even though the weight is not changed by or during capping, there are adjustments made.



Back to the earlier chart showing the instantaneous usage and the 4-hour Rolling Average. What happens when capping kicks in?

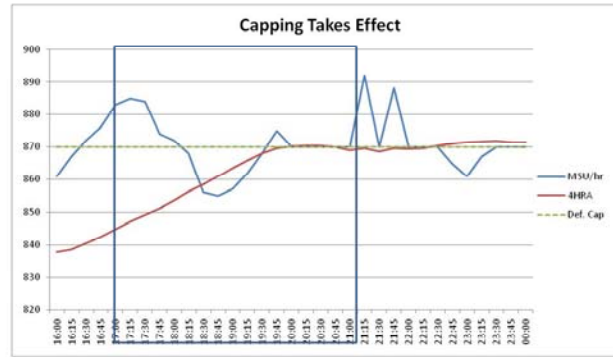


As you can see, the line gets flattened. An LPAR will not be allowed to use more than its cap value while the 4HRA is at or above the cap level. However, once the 4HRA comes down the LPAR can exceed the cap again.

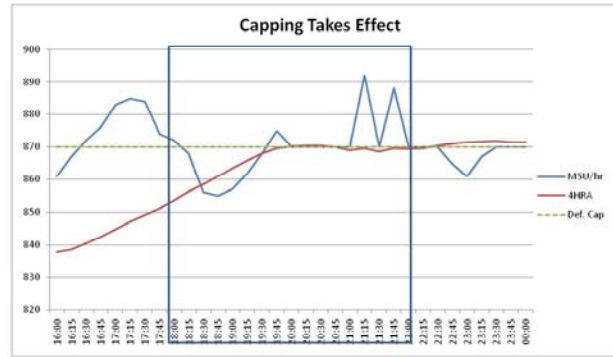


Here's a view of the 4HRA as it rolls through this time period, showing the basis for the 4HRA calculation.

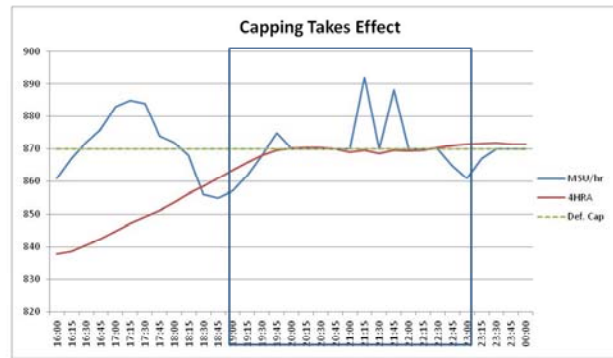
Defined Capacity in Effect



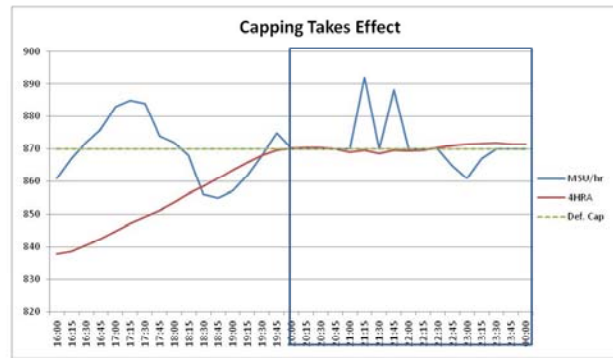
Defined Capacity in Effect



Defined Capacity in Effect



Defined Capacity in Effect



How Does Soft Capping Work with LPAR Group Limit?

- Each LPAR gets a 'Group Share' when capping occurs
- The Group Share is calculated by each LPAR independently, based on its weight and latent demand
- Each WLM is aware, in its LPAR Group calculation, of the usage of each other LPAR but not the type or importance of the work
- Decision is based on the donor-receiver model
 - If the LPAR does not need its cap level it may set a lower cap (be a donor)
 - If the LPAR has a lot of demand and some other LPAR has made cycles available it may be a receiver

Difficult idea: WLM does all this! Even though each WLM does the calculation independently, given that they all use the same calculation they all do the right thing.

The Impact of Capping

- It depends ...
- ... on your weights
 - Are they what you need for your workload goals?
- ... on the difference between your weights and the cap
 - Keep them reasonably close to avoid dramatic swings
- ... on the type of work
 - Work with long waits may miss a window
- ... on your Service Class goals
 - Goals must reflect real needs, especially when capped

The Impact of Capping

- Your weights specify what proportion of the processing capacity each partition is entitled to
- You can go over your weight while there are cycles available
- But when constrained – all partitions busy, especially when capped – the weight can have a major impact
- Question: Can you meet the goals of your more important workloads while running at your weight?

You can exceed your weight whenever there are cycles available and capping is not in effect, but as soon as other LPARs want their share you may be cut back. A low weight LPAR can give fine performance at certain times of the day and be pretty poor at other times.

You need to understand the 'necessary load' (load that must run there for some reason) on that LPAR in order to determine an appropriate weight.

The Impact of Capping

- Poorly set Service Class goals can hurt
- If your online goals are too easily attainable you may usually exceed them – PI may be as low as 0.5 - but meet them (PI 1.0) when capped
 - PI is good but service is not what the organization needs
 - The Service class may become a donor
- If your critical batch goals are too easy, especially for chains of jobs as in a Production application, you may miss due-out times while still meeting the Service Class goals

We've come across this situation a number of times when talking to installations.

The Impact of Capping

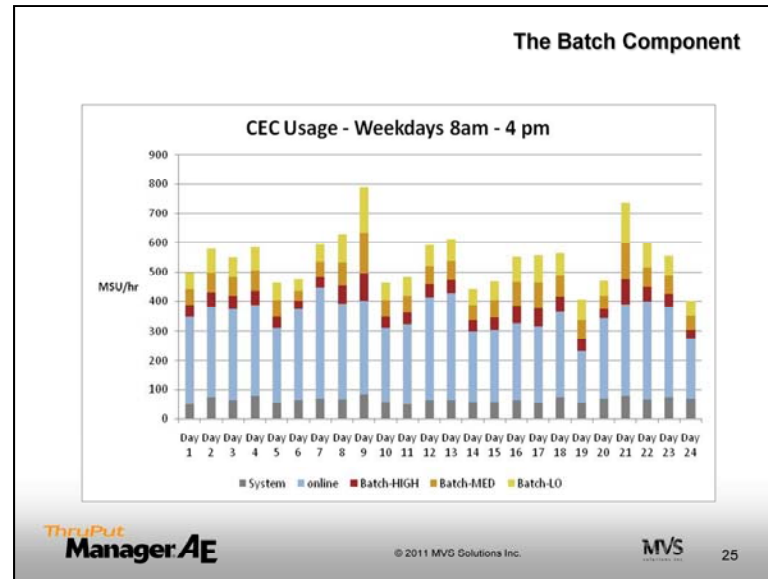
- If your Service Class goals are too hard – often get poor PI's – too much focus may be given to the highest importance work and your desired workload mix may be unattainable

The desired performance objective is to get a PI close to 1 for each workload. If there are cycles available, WLM will allow Service Classes to over-achieve, so set the Service Class goals (more correctly the Period goals) to achieve 1.0 when the system is normally busy and set the importance levels such that your 'loved ones' achieve their goals when it's approaching peak.

The Impact of Batch

- Many installations claim online drives their peaks
- In fact, most have a reasonable batch component
- JES2 initiators continue to select work no matter what
- WLM initiators slow down selection when constrained but queue length is a major factor in decisions

What's been dead for 20 years and is still haunting your datacenter? Batch



This was sent by a customer and shows the breakdown of system, online and batch cycles consumed. We've further broken the batch component down as 25% high importance, 35% medium importance and 40% low importance.

You can see that batch is a significant component of the overall load on this CEC, sometimes over 50%.

The Impact of Batch

- Batch and online on the same LPAR
- Batch and online in different LPARs on the same CEC and in the same Sysplex
- Batch and online in different LPARs on the same CEC and in different Sysplexes

Impacts differently depending on where batch is running.

The Impact of Batch

- Batch on the same LPAR as online
 - Some batch may be running in Discretionary
 - Gets a periodic slice since z/OS 1.6
 - Donors are any Service Class with a velocity of 30 or below
 - Other batch may be a receiver from any Service Class that is well exceeding its goals and can still meet its goals after donating
 - May not matter when uncapped, but ...

Quote from IBM: “But, it is also possible that a more important service class can become a donor for a less important service class, for example, if the more important service class overachieves its goals and the projections say that an adjustment will improve the less important work and still allow the more important work to easily meet its goals.”

The Impact of Batch

- Batch on the same CEC as online but in a different LPAR
 - The batch workload contributes to the 4-hour Rolling Average
 - WLM Service Class goals may help but only if within the same Sysplex
 - High batch workload on one LPAR may force the LPAR Group into capping, causing slowdowns in more important online and batch work
 - WLM is unaware of the importance of a load in another Sysplex
- Example: one site filled up a batch LPAR on a test Sysplex with discretionary work – caused the 4HRA to be very high and cost them \$\$\$\$\$

Service Classes have a context of the Sysplex but apply at the LPAR. If the LPAR has cycles WLM will allow the batch work to overachieve.

The WLM instance on an LPAR in another Sysplex is unaware of the fact that it's low importance batch load is impacting high-importance online in another LPAR.

Controlling Your Costs

- Cost control must be balanced with meeting the needs of the organization
- Prerequisite 1: set proper weights that reflect the importance, urgency and volume of the workloads
- Prerequisite 2: make sure your Service Class goals and importance levels match organizational needs

WLM cannot manage your workloads well when constrained if these are poorly set

Rule 1: don't shoot yourself in the foot.

The real goal – unless senior management tell you otherwise – is to meet the needs of the organization.

Saving money while not meeting the goals is a false economy – it may cost the business far more than you save.

Controlling Your Costs

- Can you control your onlines?
- Basically No!
- Typically critical to your organization
- Highly visible to senior management
- The only control is the response time you set in your Service Class goals
- You cannot limit the volume of transactions
- Could be career-limiting to try

Onlines, other than in exceptional circumstances, are untouchable. Don't try, other than setting response time goals that make sense.

Controlling Your Costs

- Can you control your production batch?
- Yes, to a limited degree
- Some provides the foundation for your onlines
- Some is highly visible to senior management
- You cannot limit the volume of data
- But, not all production batch is equal
 - Reports compared to database updates
 - Externally focused compared to internally focused
 - Government regulated compared to unregulated
 - Penalties

Production batch is not untouchable. If in doubt look at your DR plan – which batch must get done?

Controlling Your Costs

- Can you control your non-critical batch, production or otherwise?
- Yes!
- But not by using Discretionary!
 - With Discretionary you give up control
 - Remember, it can be a receiver from any work with a velocity of 30 or less
 - Better to set a Service Class goal with low importance and velocity
- If you must use Discretionary use a Resource Group with a maximum

The only real leeway you have, other than ensuring your weights and Service Class period goals make sense, is with medium and low importance batch. However, that's often a significant load.

Our Contention

- We contend:
- *You can control the 4-hour rolling average for a CEC by controlling the selection of batch and selectively adjusting the Service Class of lower importance work that is currently executing*

Our Contention

- With ThruPut Manager AE, you set a capacity limit
 - Defined Capacity
 - LPAR Group Limit or
 - AE “Target capacity”
 - Used when no capping is desired for non-batch workloads
 - Only restricts batch workload when target is neared
- You specify up to five percentage levels of that capacity at which you want to take one or more actions
 - Limit a category of batch to n concurrent jobs,
 - Stop selecting a category of batch,
 - Change the Service Class of a category of batch to one with Discretionary and a Resource Group maximum until things get better

ThruPut
Manager AE

© 2011 MVS Solutions Inc.

MVS

34

Some people are afraid of capping because of the impact it can have on critical batch and online work. AE provides its target capacity so that you can control your batch without impacting your critical workloads.

Our Contention

- Can we prove our contention? Not yet.
 - Recent release
 - Customers in the process of implementing
 - Hope to have customers prove it by summer and talk about it at Share in Orlando

Summary

- Sub-capacity pricing can save significant money on mostly IBM software
- You must be careful!
- Understand your organizational needs
- Review your weights and Service Class goals
- Batch control is the key