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.
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.
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.
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.
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.
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

Capping Takes Effect

- Name1
- Name2
- Def Cap
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
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.
We’ve come across this situation a number of times when talking to installations.

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
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.
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%.
Impacts differently depending on where batch is running.
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.”
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.
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.
Onlines, other than in exceptional circumstances, are untouchable. Don’t try, other than setting response time goals that make sense.
Production batch is not untouchable. If in doubt look at your DR plan – which batch must get done?
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
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