

Tailored Fit Pricing: How To Manage Workloads in a World Without Capping

Nathan Brice
IBM

November 2019
Session 00



Important Disclaimer

IBM's statements regarding its plans, directions and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

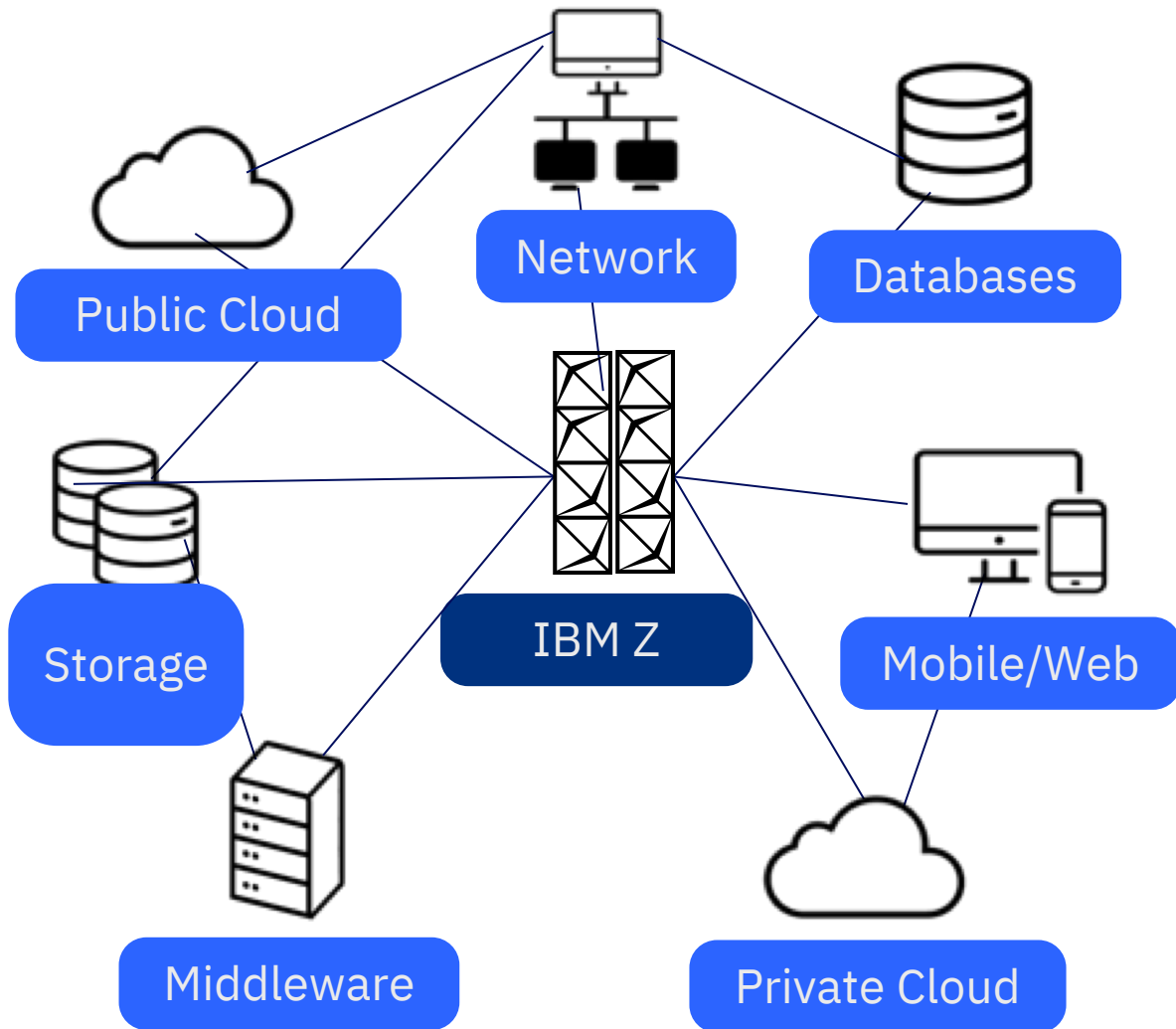
The information in this document is provided "as is" without any warranty, express or implied, including without any warranties of merchantability, fitness for a particular purpose and any warranty or condition of non-infringement. IBM products are warranted according to the terms and conditions of the agreements under which they are provided. It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs.

- **Performance.** Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.
- **Customer Examples.** All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.
- **Availability.** Not all offerings are available in every country in which IBM operates. This document is current as of the initial date of publication and may be changed by IBM at any time.
- **Trademarks.** IBM and the IBM logo are trademarks of International Business Machines Corporation, registered in many jurisdictions. Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates. Other company, product and service names may be trademarks, registered marks or service marks of their respective owners.

Highlights

- Tailored Fit Pricing is a revolutionary new pricing model that eliminates the need for workload capping and provides a complete alternative to the Rolling 4 Hour Average
- Transitioning to a model where measurement of overall MSU consumption is critical represents more than just a change in how you pay for software on IBM Z. Several stakeholders are affected and various considerations around how workload is managed must be made
- Effective use of tooling can help you: IBM Z Decision Support for Capacity Planning delivers pre-defined dashboards to give visibility in to current consumption levels, forecast future consumption and provide insight into where resource optimization can take place

What is Tailored Fit Pricing?



Digital Transformation is leading to Hybrid Cloud pattern of deployment across enterprises

IBM Z is a critical part of this infrastructure

Unpredictable demand in era of Hybrid Cloud



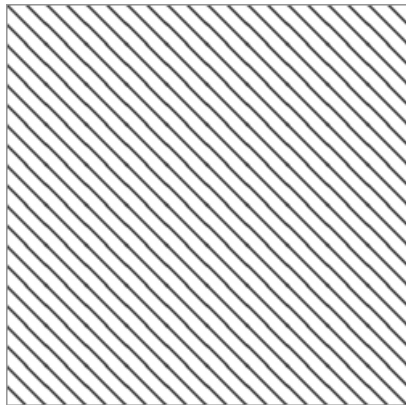
In the era of hybrid cloud, where everything is connected and workload patterns are constantly changing, predicting demand for IT services can be a major challenge.

Evolution of Z Software Pricing

1970 - 1999

PAST

Full Capacity



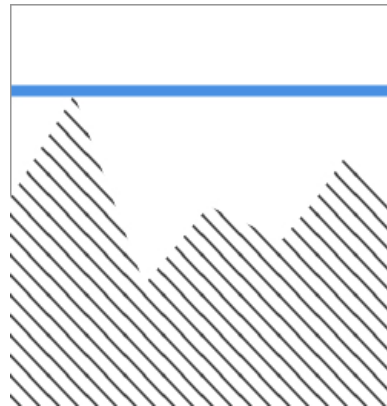
- Simple way to charge for z/OS-based software

1999 - 2019

PRESENT

Sub-Capacity

(R4HA)

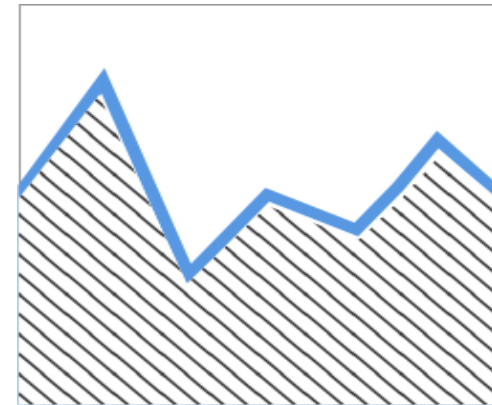


- Modelled on 90% utilization
- As system size increases, align product value to less than full capacity

2019 onwards

FUTURE

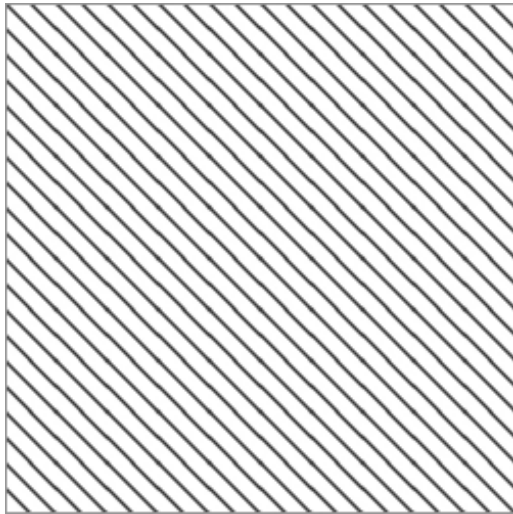
Tailored Fit Pricing



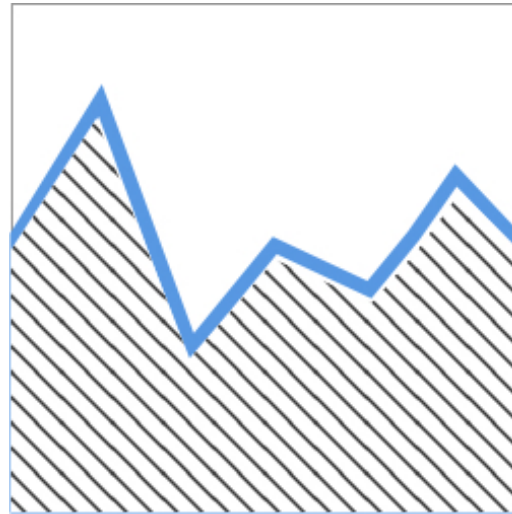
- Remove R4HA
- Align value to the workload for the amount of system resources it consumes

Models of Tailored Fit Pricing

Enterprise Capacity



Enterprise Consumption



An MSU consumption pricing model that allows clients to:

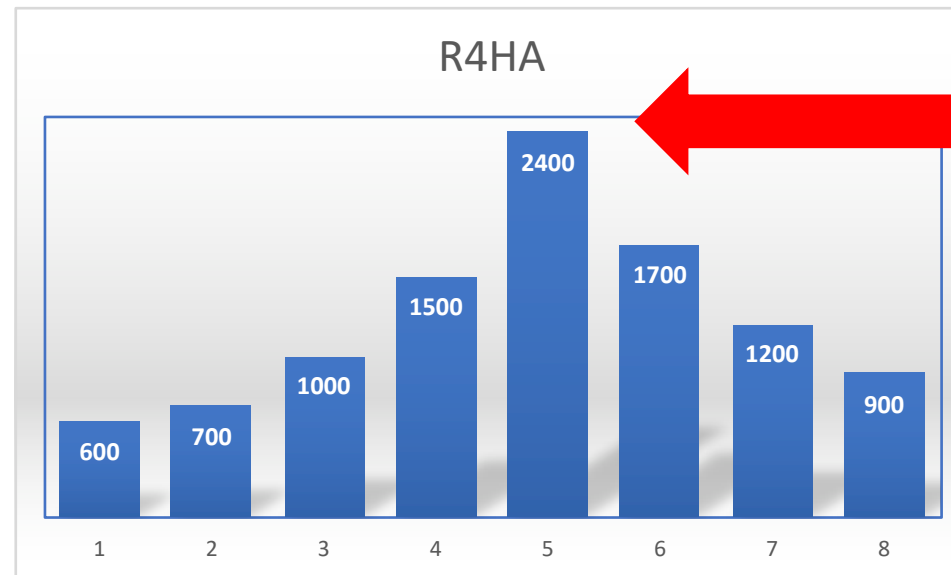
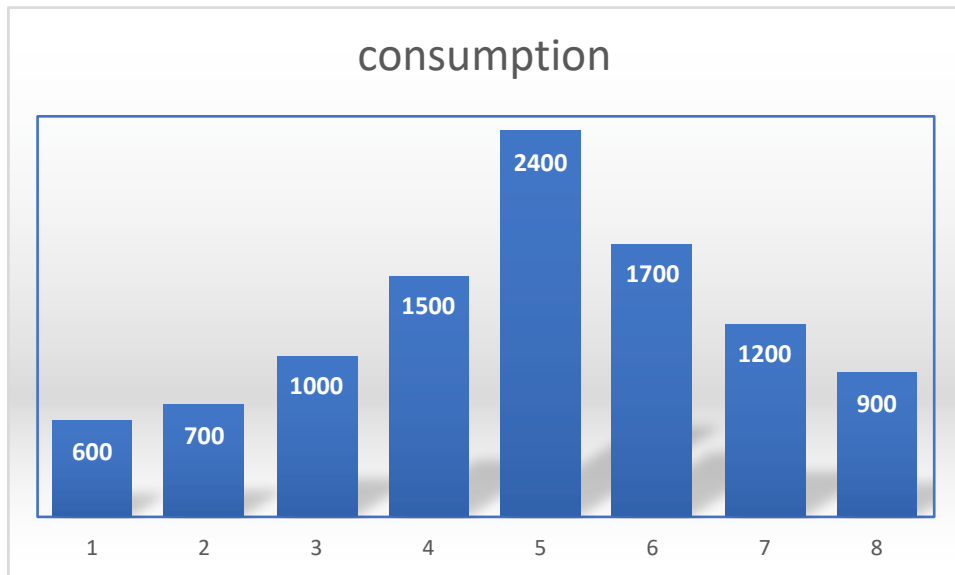
- Take full advantage of the hardware they own
- Peak and spike without 'penalty'
- Smooth seasonal variations over the entire year
- Grow at a highly competitive per MSU price
- Pay for workloads with price consistency

We're going to focus more on the Enterprise Consumption model today

Overall Usage Determines Charges, Not Peaks

Peak and spike as the business demands...without blowing the budget

- In both workloads below, the total MSUs consumed is 10,000 over the same period of time
- Under R4HA, the single spike of 2,400 MSUs sets the price for the entire month
- **Under consumption, you pay for exactly what you use at the same rate**



This determined
the price under
R4HA

What exactly is MSU consumption?

Capacity reference (capacity markers to measure entitlement)

- **Full-cap** = size of the machine, based on HW MSU ratings; Sub Capacity Reporting Tool (SCRT) not required
- **Sub-cap** = combination of R4HA peak, LPARs where run, metric type, etc., as reported by E5/B5
- **Total MSUs** as reported by K5 or N7 sections

In SCRT:

Sub-cap:

PRODUCT SUMMARY INFORMATION		
MLC Product Name	MLC Product	Tool MSUs (
z/OS V2	5650-ZOS	1098
DB2 11 for z/OS	5615-DB2	1098
CICS TS for z/OS V5	5655-Y04	1098
IBM MQ for z/OS V8	5655-W97	1098
OPC V2	5697-OPC	1098

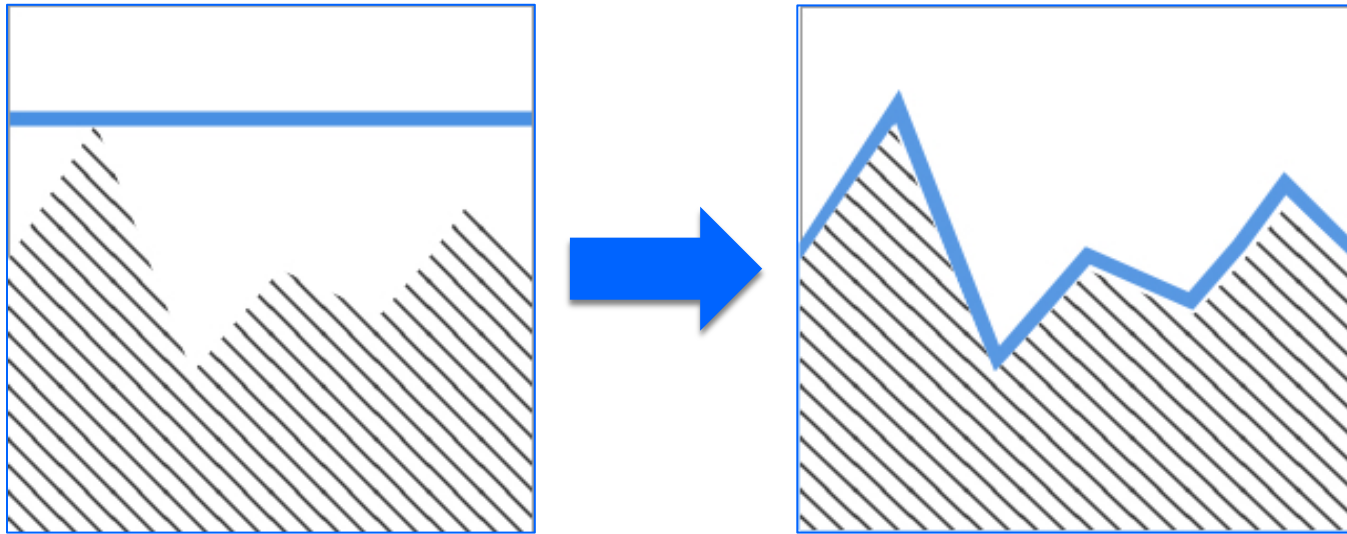
Consumption:

DETAIL LPAR USAGE DATA SECTION		
	Total MSU Cons	Peak Hour I
SYSC	304969	851
SYSE	163251	631
CPC	468220	1189

ACTIVE CONTAINERS	
SCRT Container Identifier	Solution ID
CPS1	Z194E15-F44853D-E56

==CPS1==	
Solution ID	Z194E15-F44853D-E56
Solution Name	Production Container
Peak Four Hour Rolling Average	1098
Total MSU Consumption	468220

Defining the MSU Baseline



The baseline should account for your previous MLC and IPLA usage. Growth beyond that is calculated at a reduced rate

Baseline established upon 12 months of SCRT reporting

This should allow seasonal variations to be accounted for plus considerations for growth

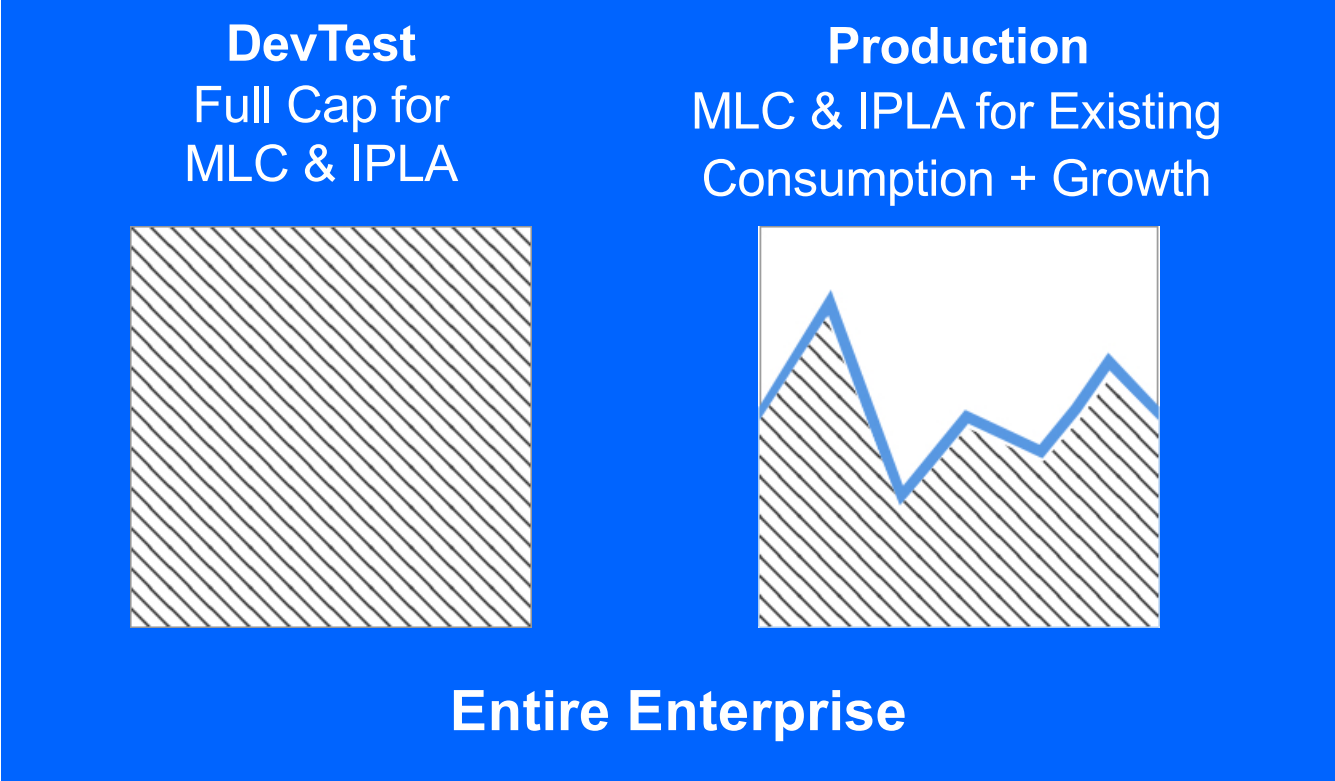
Therefore important to understand what workloads you have running within your environment to start and allocate to the container

Typical Enterprise Consumption Solution

Before



After



What are the challenges today for
managing **cost** and **performance**?

Key User Personas



Gemma
CIO

- Finds it difficult to figure cost and believes IBM Z is too expensive compared to other platforms
- Thinks there is a shortage of IBM Z skills and talent but knows her business is dependent on quality of service levels



Dan
IT Architect

- Configures system to handle peaks
- Projects technology trends to make choices and depends on performance testing understand impact of changes
- Doesn't know how much an application costs



Carl
Capacity Planner

- Has trouble forecasting demand for next 3-4 years
- Has to reduce service towards the end of the month
- Spends too much time on admin, tuning, instead of bringing on new workload

Under the R4HA...

Controlling cost is vitally important. We don't want any surprises on operational cost so we can accurately manage our budgets



Gemma
CIO

I work to ensure the applications that must run within the peak are optimized but I don't have time to look at anything outside



Dan
IT Architect

My focus is on ensuring the peak period is kept in control, including capping and planning long term to manage operational costs



Carl
Capacity Planner

...there are consequences



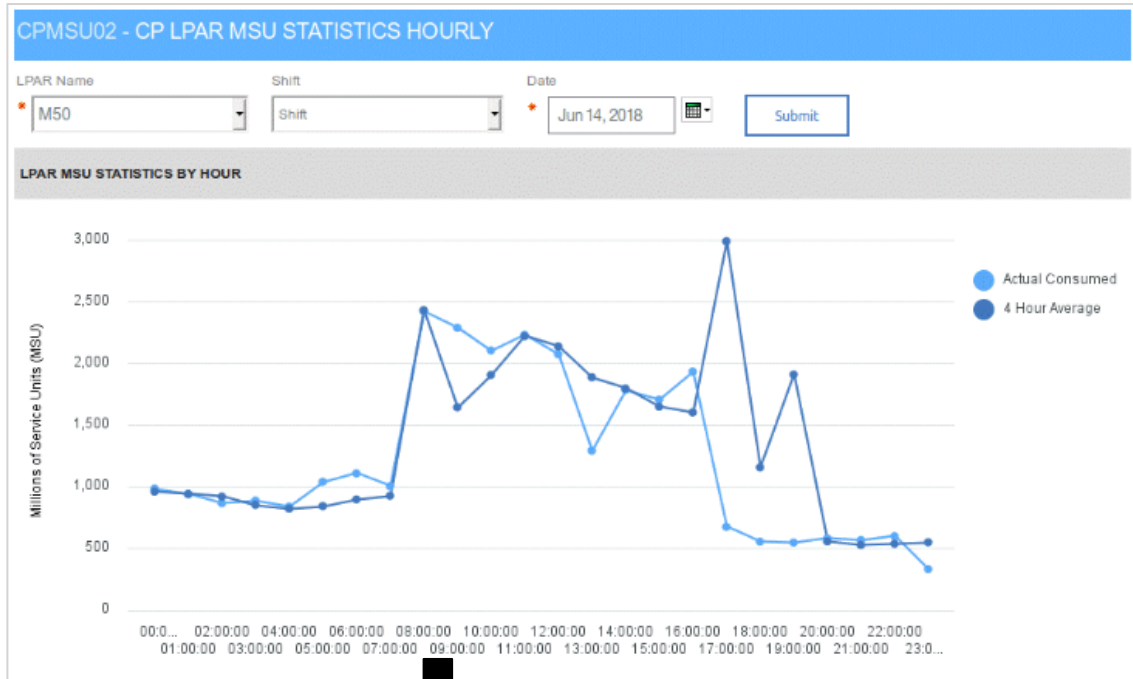
I use several tools to track costs within the R4HA peak and implement various capping strategies.

It's easier for me to understand the costs on our public cloud contract. We should deploy new work there

My applications are affected by our capping processes. Sometimes our SLAs are impacted and workloads are not taking full advantage of the capabilities of our mainframe

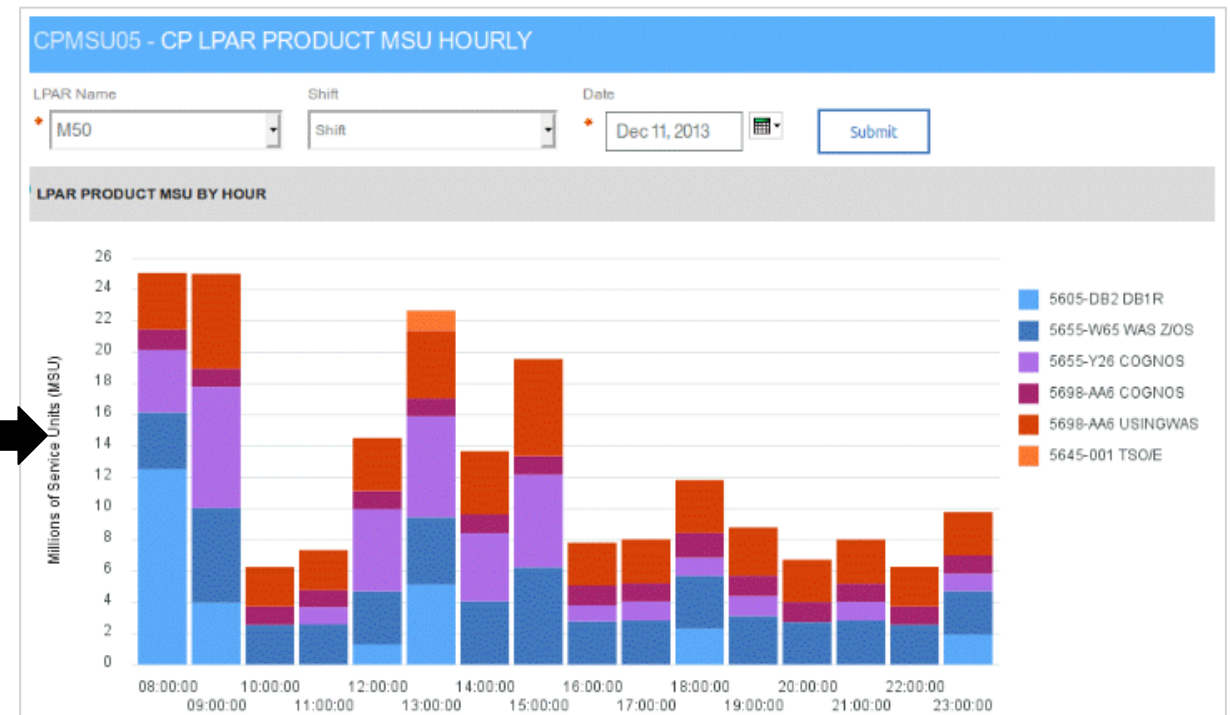


R4HA Reports



Identifying the peaks earlier can avoid surprises prior to SCRT reports are generated...

...including drill down to a product's contribution to the R4HA



Capping can negatively impact workload



- System outages
 - Resources not being freed in timely fashion
 - Storage shortages
 - Work (e.g. Service Request Blocks (SRB)) backed up, common storage shortage
- Important work displaced or Service levels missed
- Less important work displaced
- Increased response times or CPU delays

Capping in a "Consumption based" installation ?

We used to cap workloads to limit our costs. Is there any benefit to this now we've switched to an MSU consumption model?

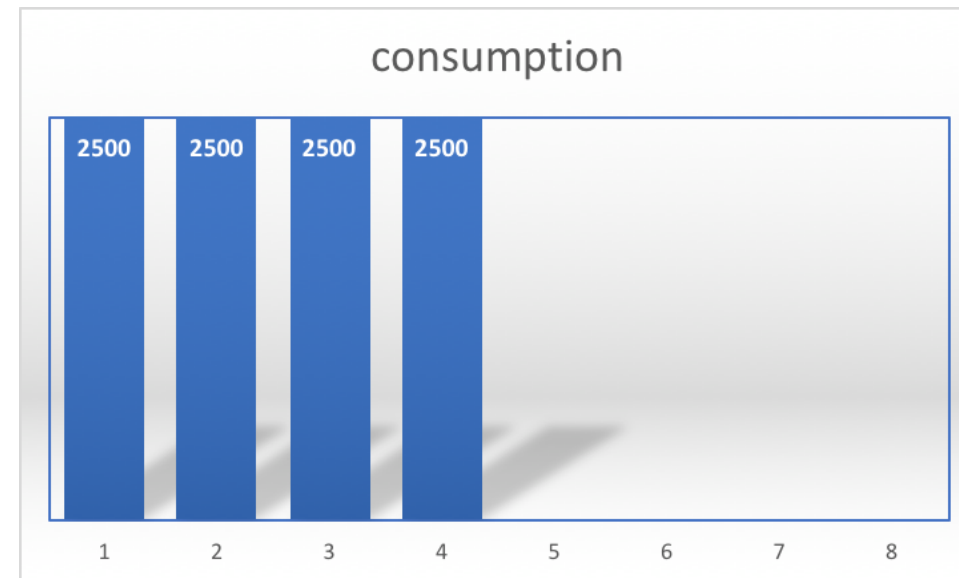
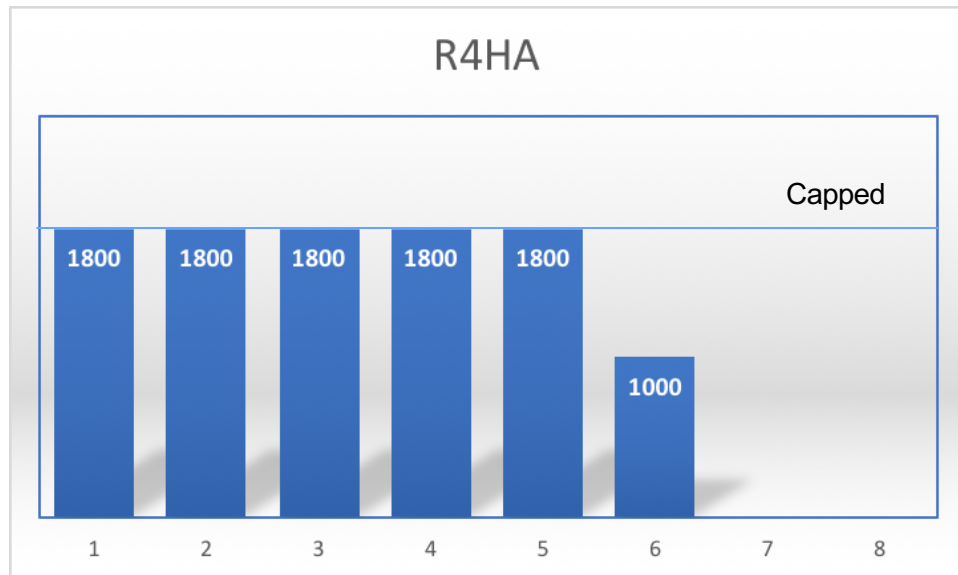


- Everything running in one specific 4 hour window each month impacted the cost. Outside the window you could perceive the workload as “free”
- The various capping algorithms helped clients to limit the MSU *peak* usage – with all negative consequences described earlier.
- In a "Consumption based" installation capping is irrelevant – as the pricing is not derived from the peak, but rather from every MSU consumed. Thus, controlling peaks alone (as capping does) is not helpful.
- As there can't be a "control via capping" in a "Consumption based" installation – other methodologies need to be applied.

What if we removed capping?

Maximize the hardware, minimize the batch window:

- Let's assume the nightly batch requires a total of 10,000 MSUs to complete
- Let's assume the machine is rated at 2,500 MSUs, but capped at 1,800 for the R4HA
- **By removing unnecessary soft caps, batch windows can be dramatically reduced**



How does transitioning to
Tailored Fit Pricing affect **you**?

Under Tailored Fit Pricing new challenges await

We need to understand the level of our current MSU consumption & future use so we can understand impacts of workload changes



My applications are being driven more by external sources. Without capping how do we ensure we are not impacted by spiky workloads



I need to assist in projecting the needs for growing workloads on the mainframe and fit in with our MSU allocation



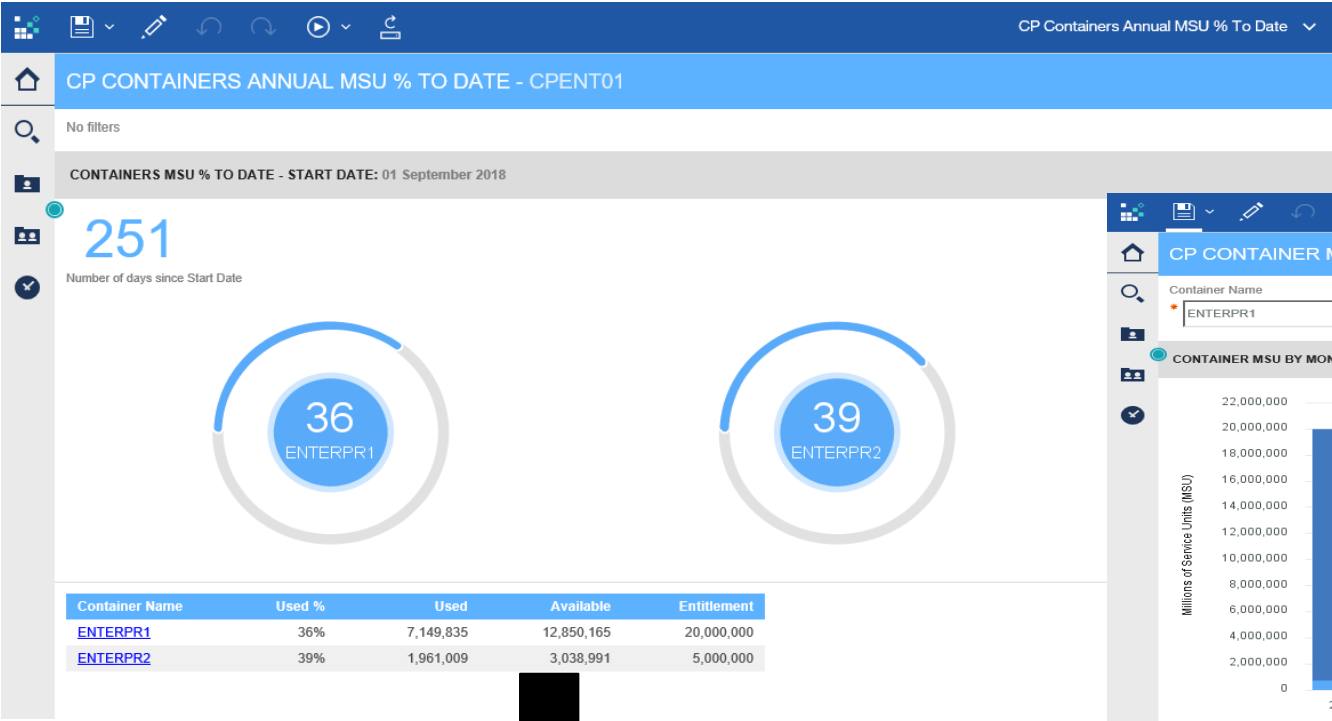
Challenge #1: Measuring (and forecasting) MSU consumption

We need to understand the level of our current MSU consumption & future use so we can understand impacts of workload changes

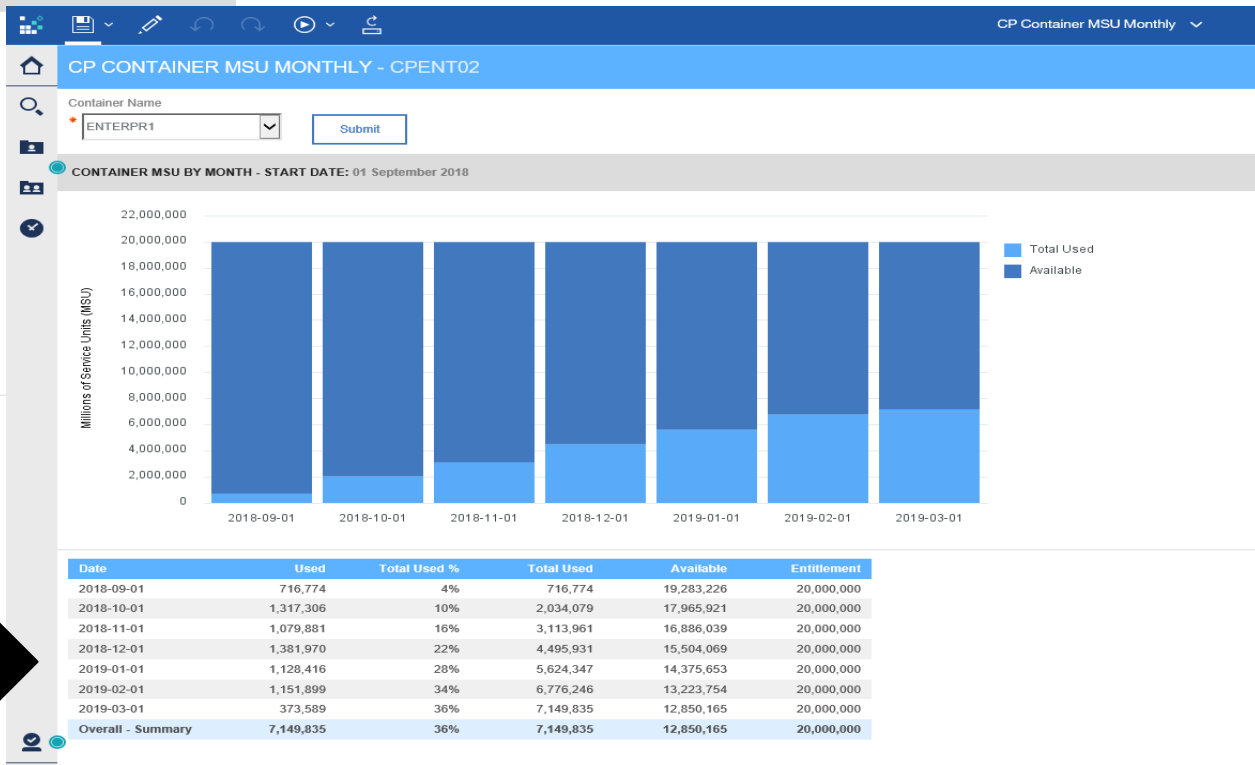


- The transparency of consumption based pricing is attractive to many customers
- Forecasting future consumption can help with quantifying costs, tie back to business decisions and avoid surprises
- When forecasting knowing that you may exceed your allocated baseline can be a good thing (or maybe a bad thing!)

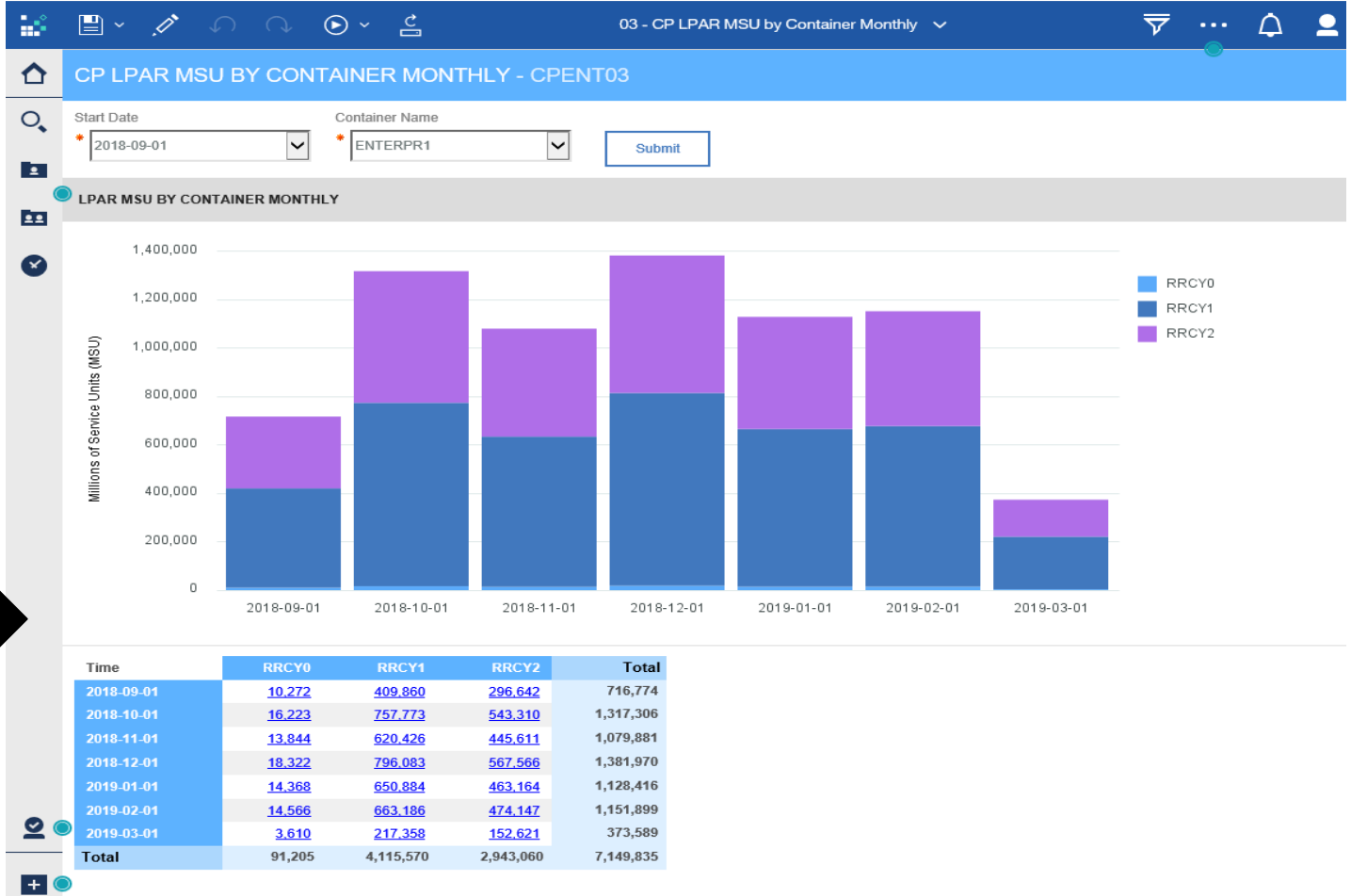
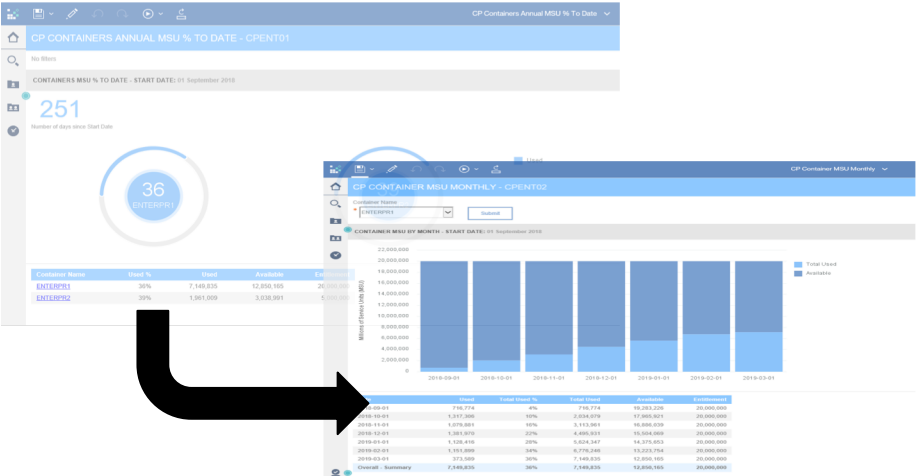
The need for clear views of the current state



Overall consumption so far versus annual projections

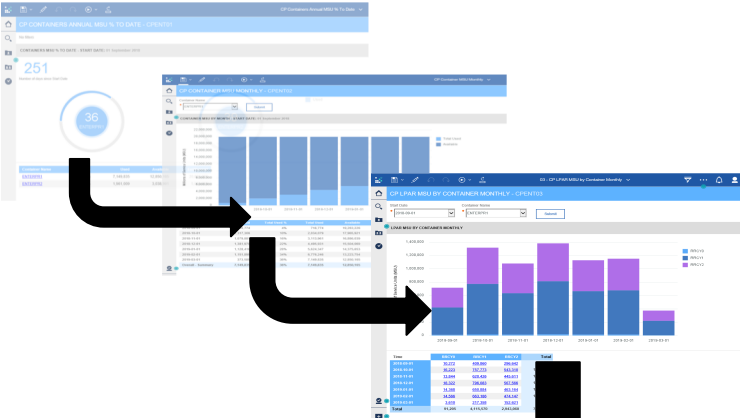


Where the workload is running

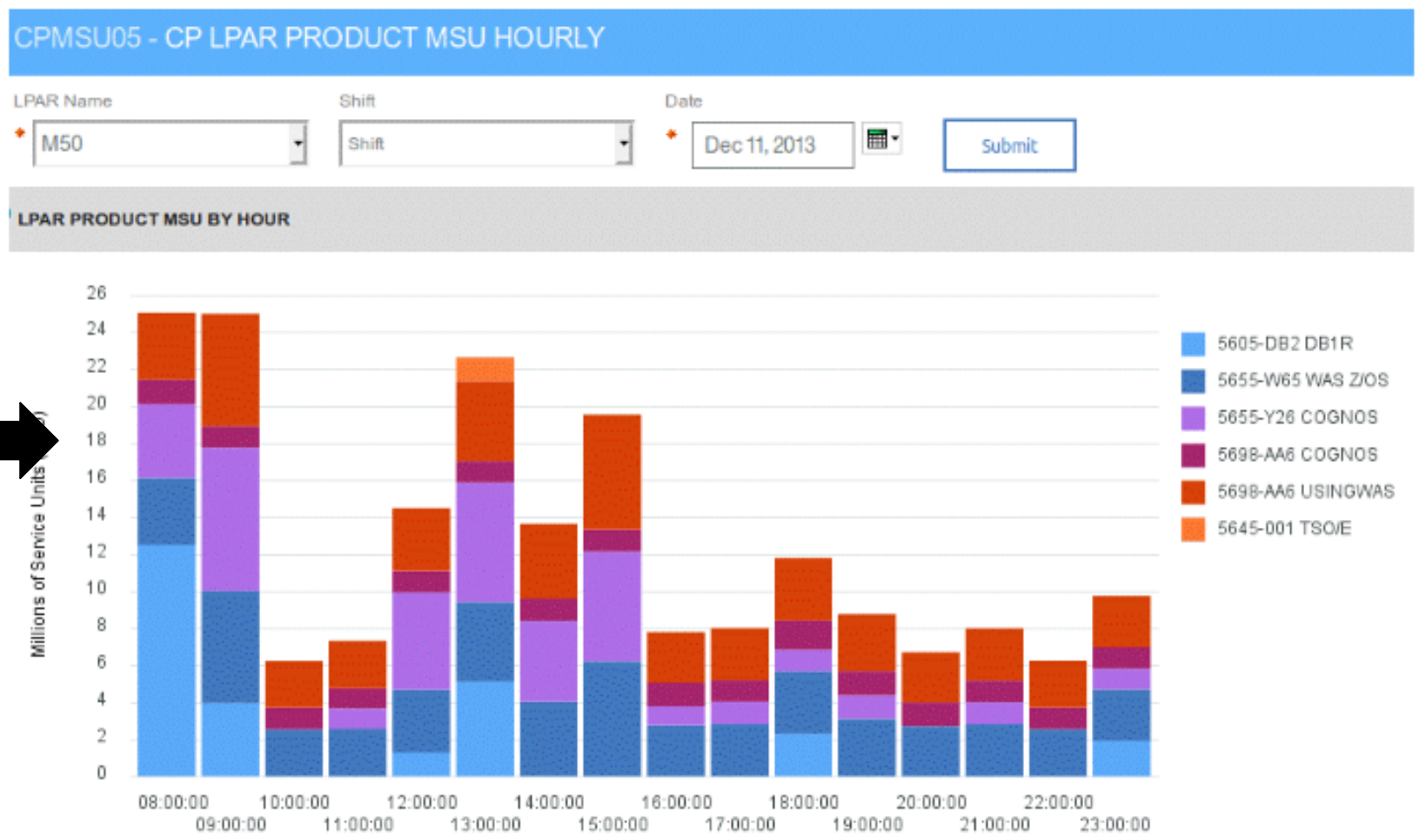


Now can see on an LPAR or Service Class basis what workload was running and how much did it consume

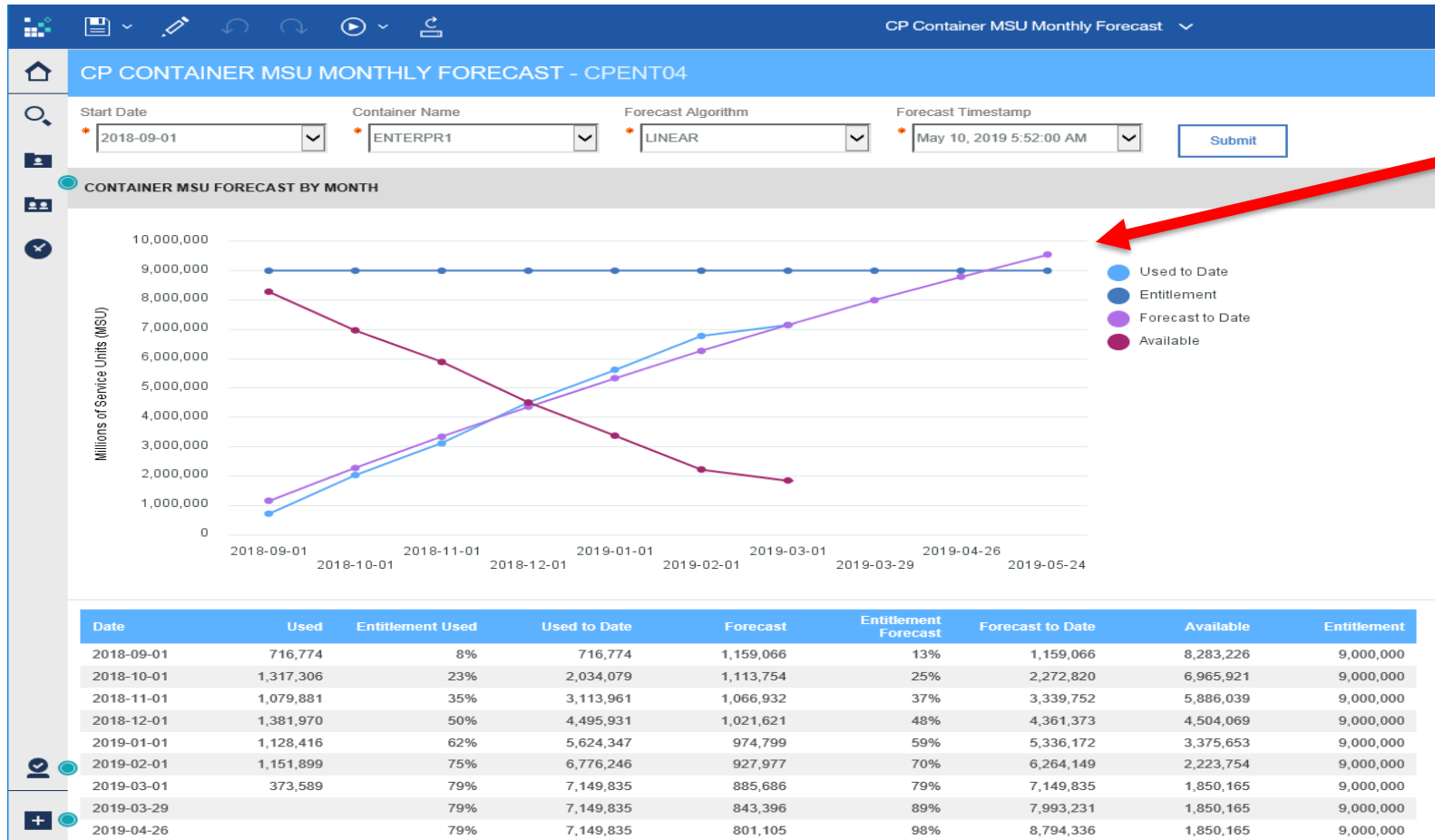
And which workloads are contributing



Drilling down further into the jobs or products that are consuming MSUs



Forecasting is critical to avoid surprises...



The forecast report indicates when we believe MSU consumption will exceed the baseline for the current year

It is for the business to decide if this is a good or bad thing! The key is that you have the visibility to make an informed decision

...and tie costs back to the business

We're being asked by the business to support new sales events, can we estimate the impact on workloads?



We know what applications will be affected and can see the workload growth so costs can be identified and incorporated into our chargeback process



I can tie this back into my capacity planning dashboards to know that we can support these changes

Challenge #2: Ensuring workloads can be managed effectively

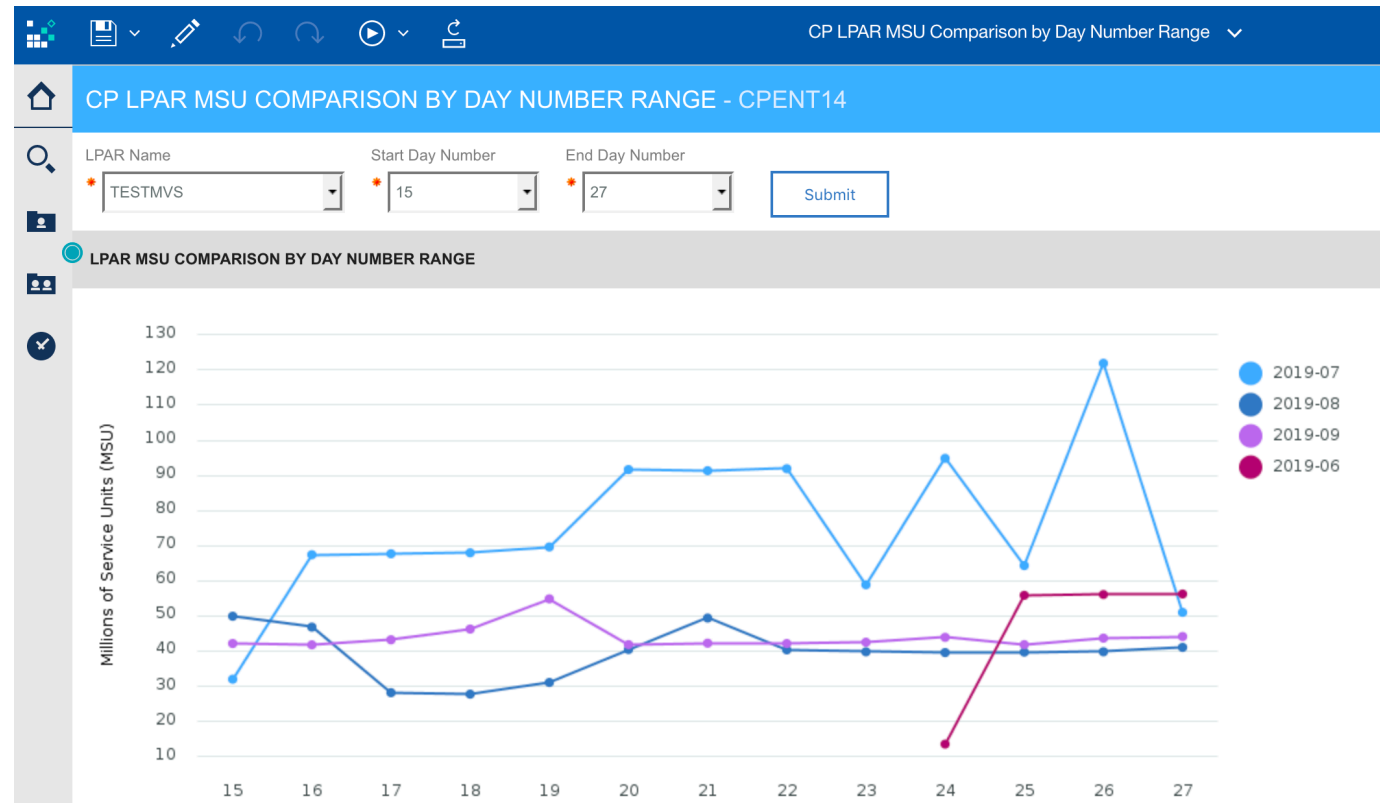
My applications are being driven more by external sources. Without capping how do we ensure we are not impacted by spiky workloads



- Capping gave customers a feeling of reassurance and removing it to exploit the full value of the hardware BUT concern can be that consumption can run away and end up costing more
- Existing tooling focused to R4HA cost control or based of SCRT / RMF report may not have provide the right level of granularity here

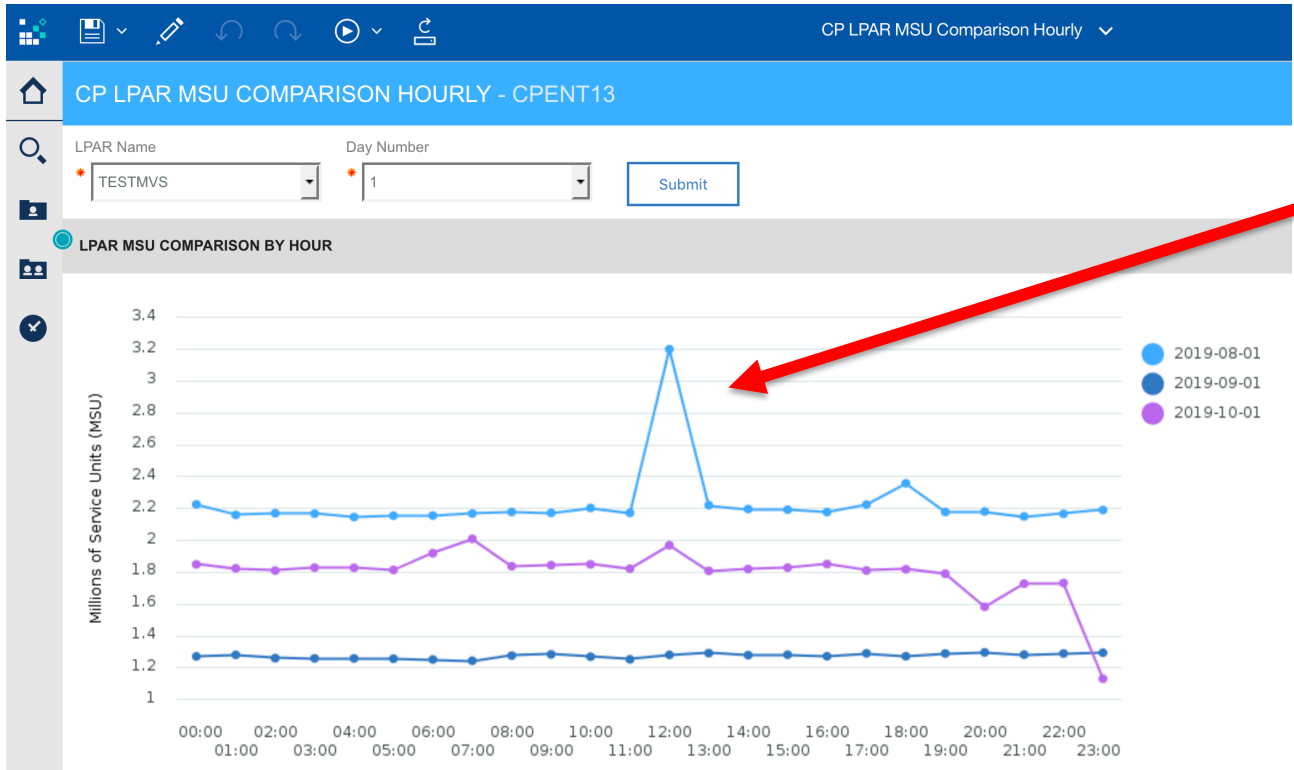
Actions to track changes in consumption

- Comparing past days with actual days can help to understand if the consumption is within expectation or not.
- For example: Comparing the same final days of each month, or maybe days of the week across the month to know when typical peaks occur
- Using this benchmarking can help in setting understanding of application performance. Comparing “wrong” days can lead to incorrect assumptions



Day to day comparison of workload MSU and drill down to hourly levels

Actions to track changes in consumption



An obvious spike occurs on one LPAR during a given hour. This can be investigated through drill-down

Ensure the right alerts are in place

Our monitoring and other tooling has been tuned to alert and throttle based on experience with the R4HA.



When our application changes are applied, the behavior may change and drive up MSU. I need to consider the impacts



Check what alerting and exceptions you are checking for today

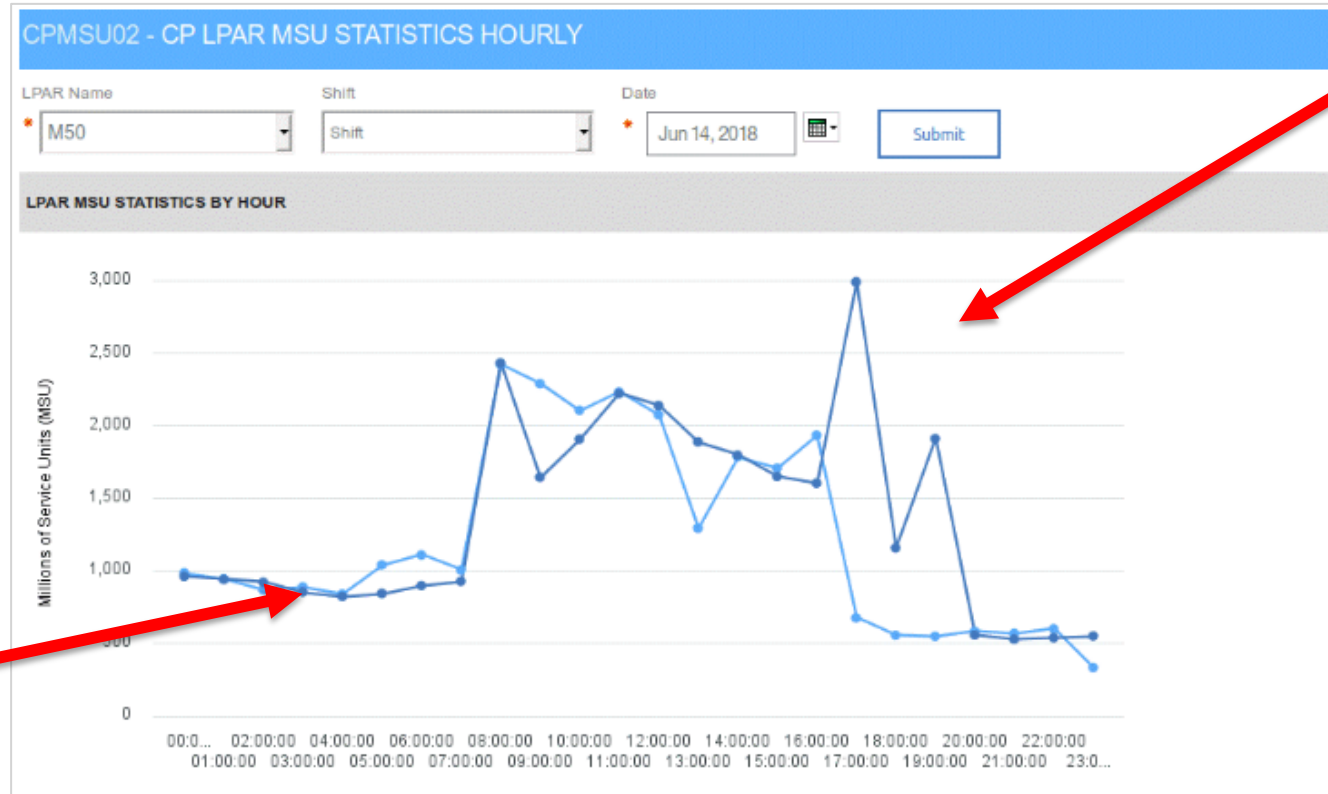
Challenge #3: Optimize use of MSU allocation

I need to assist in projecting the needs for growing workloads on the mainframe and fit in with our MSU allocation



- Note we are focusing now on workload 24/7/365, not just a 4-hour peak period!
- There are many options we take advantage of here:
 - Are we making effective use of zIIP capacity?
 - Are older applications running efficiently exploiting hardware and compiler updates?
 - Would we benefit from a database reorg or looking at our backup policy?
 - Health check of subsystems and looking at resources
- A holistic approach will identify the areas ripe for optimization

Need to focus on ALL workloads

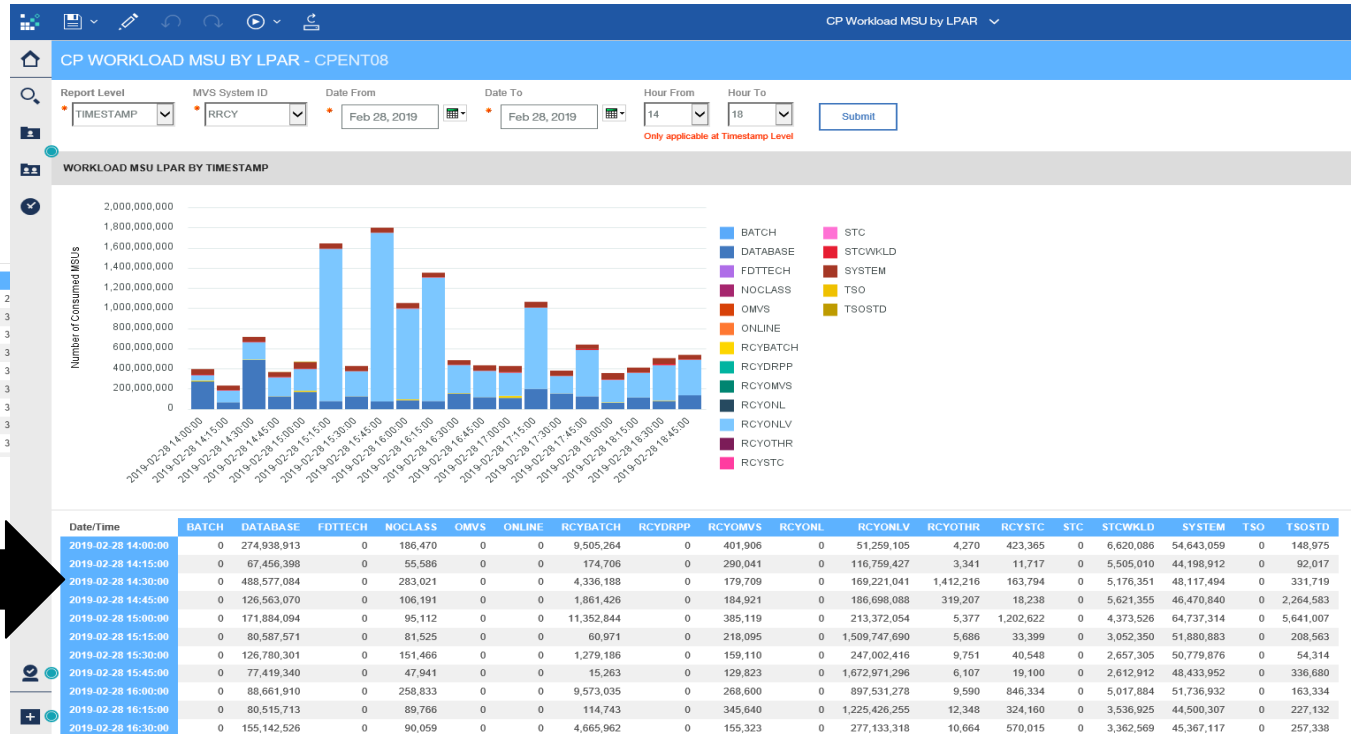
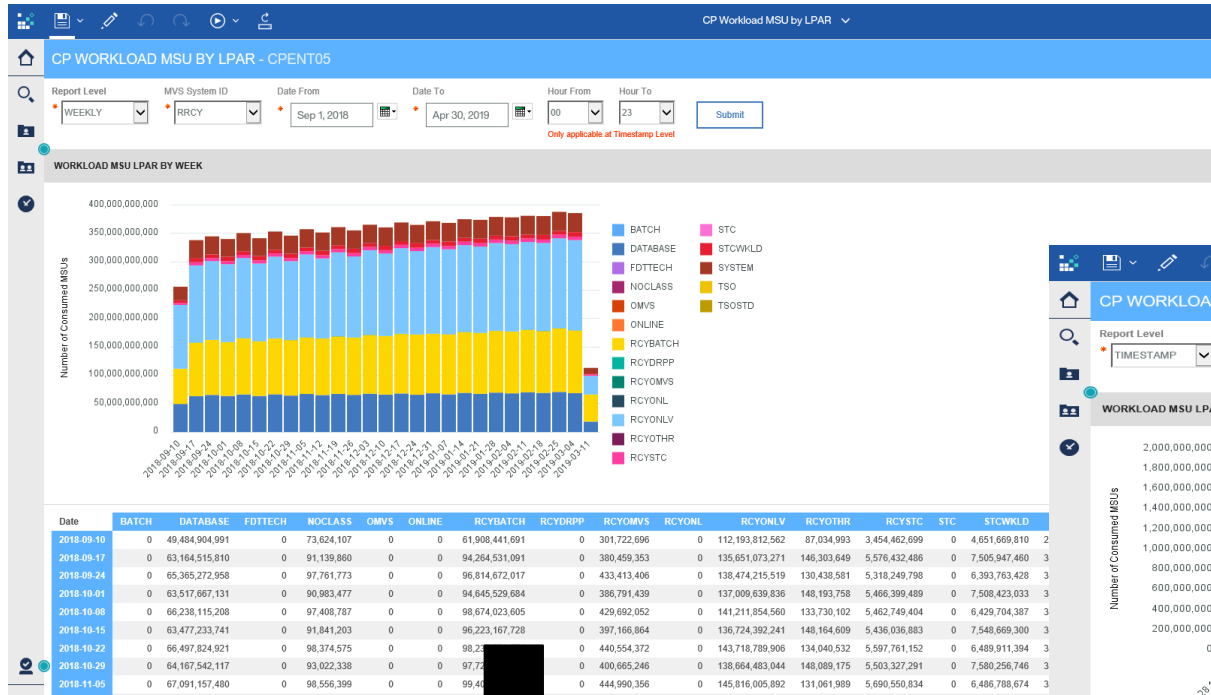


Previously, focus would be on optimizing workload within the 4 hour peak

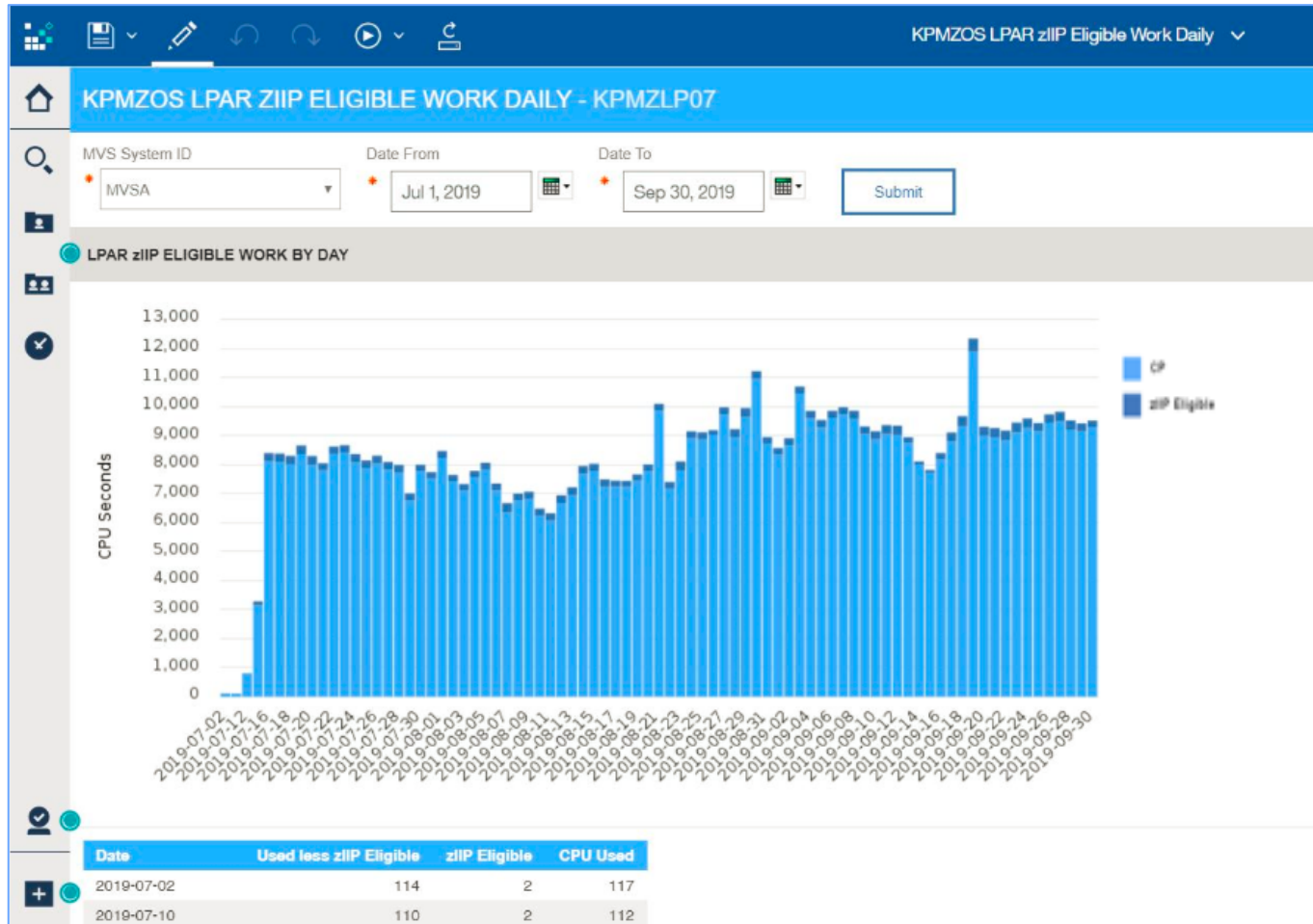
Now workload outside of the peak period is just as important to manage

Drill-down to identify the highest consumers

MSU Consumption by Workload and Service Class Details on the consumption based on daily / hourly / timestamp granularity



Look for opportunity to reduce MSU usage



Understand levels of zIIP offload

What solutions are **available**
today to support you?

Supporting Your Journey to Tailored Fit Pricing

Planning

Understanding your current applications and workloads including which products are running on each LPAR

Key tools:

- Tivoli Asset Discovery
- IBM Z Decision Support for Capacity Planning

Managing

Ongoing analysis to visualize current MSU consumption levels per container and forecast future consumption

Key tools:

- **IBM Z Decision Support for Capacity Planning**
- IBM OMEGAMON

Optimizing

Maximizing the value of your deployments to drive efficiency and performance improvement

Key tools:

- Compilers
- Specialist tools (for example: IMS Buffer Pool Analyzer, IBM Z Batch Resiliency)
- IBM Application Discovery and Delivery Intelligence

Managing with IBM Z Decision Support for Capacity Planning

Announced September 2019: ibm.biz/IZOIAnnounce



IBM Z Operations Insight Suite

IBM Z Operations Analytics

Data Streaming, Problem Identification & Anomaly Prediction

IBM Z Decision Support for Capacity Planning

Data Streaming, Performance Analysis, Capacity Forecasting, Cost Management

Visualize on the platform of your choice

IT Operational Analytics platforms



Multiple Stakeholders – Single Source of Truth



Z IT Operator



IT Operations Manager



Z Capacity Planner

IBM Z Operations Insight Suite

Data Streaming

Data where and when you need it

Problem Identification

Rapid operational root cause analysis

Anomaly Prediction

Prevent outages with Machine Learning

Performance Analysis

Gain insight for critical decisions

Capacity Forecasting

Predictive resource usage & optimization

Cost Management

Optimize Tailored Fit Pricing strategies & enable chargeback



Z SME



Line of Business Owner



Head of Mainframe

Learn More

IBM Z Operations Insight Suite on IBM Marketplace

The latest updates and information about IBM's leading solution for managing IBM Z as part of a hybrid cloud environment

ibm.biz/IZOInfo

ibm.biz/IZDSCapPlanInfo

Tailored Fit Pricing: How to manage workload in a world without capping

Discover how IBM Z Decision Support for Capacity Planning can help manage, forecast and optimize your MSU consumption

ibm.biz/IZDSCPandTFP

Achieve Operational Excellence

Learn how to optimize complex hybrid cloud environments with efficient systems management and operational insights

ibm.biz/OpExcellence

IBM Z Software Newsletter: Operations & Management Edition

Subscribe to the quarterly newsletter for operation, systems programmers and administrator to get the latest news, tips, blogs and trials in one place

ibm.biz/ZOperations

Summary

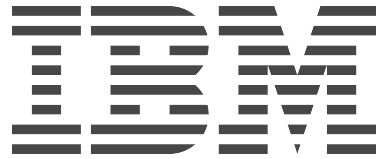
Tailored Fit Pricing is revolutionary new pricing model for software on IBM Z that eliminates the need for capping and provides a complete alternative to the R4HA.

To be successful requires a change in how you look at managing workloads. Instead of capping to contain costs, optimizing existing applications is the key and this cannot be done without a clear view on current consumption levels

With a clear view on usage levels of workloads, accurate forecasts can be made that tie back to business drivers and your capacity planners are empowered to model workload growth on the mainframe

IBM Z Decision Support for Capacity Planning, as part of the IBM Z Operations Insight Suite, gives unique visibility into current and future MSU consumption and provide guidance into how to optimize workloads to control costs

Reach out to us for a deep-dive discussion on how you can leverage the current capabilities or join one of our Proof of Technology workshops



Please submit your session feedback!

- Do it online at <http://conferences.gse.org.uk/2019/feedback/oo>
- This session is OO

1. What is your conference registration number?

 This is the three digit number on the bottom of your delegate badge

2. Was the length of this presentation correct?

 1 to 4 = "Too Short" 5 = "OK" 6-9 = "Too Long"


1 2 3 4 5 6 7 8 9

3. Did this presentation meet your requirements?

 1 to 4 = "No" 5 = "OK" 6-9 = "Yes"

1 2 3 4 5 6 7 8 9

4. Was the session content what you expected?

 1 to 4 = "No" 5 = "OK" 6-9 = "Yes"

1 2 3 4 5 6 7 8 9