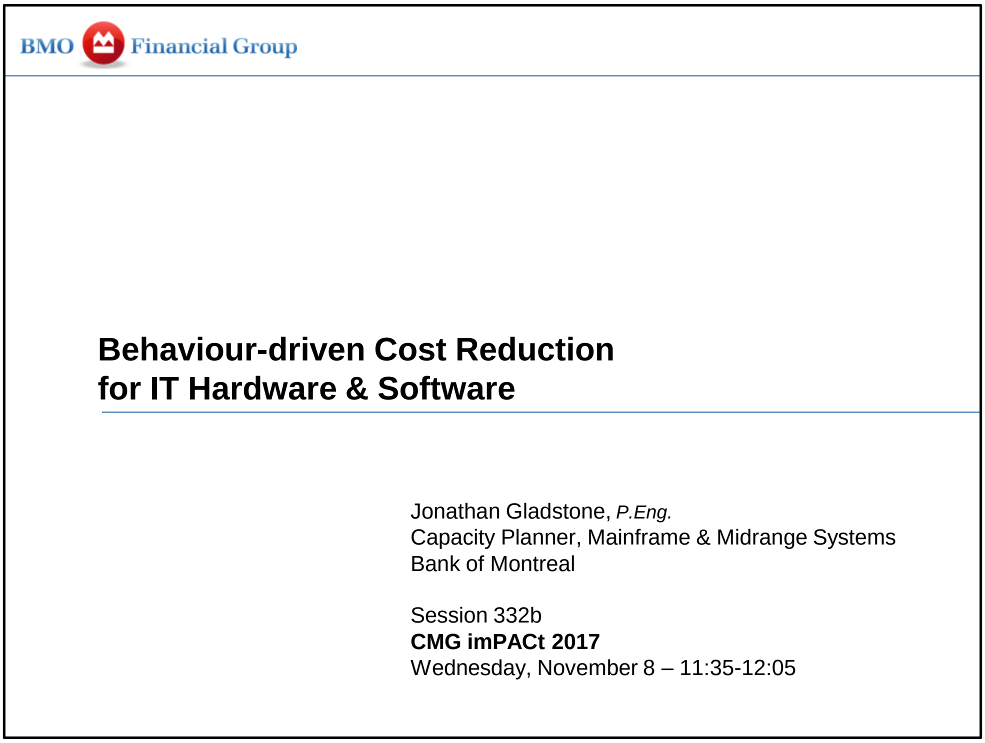
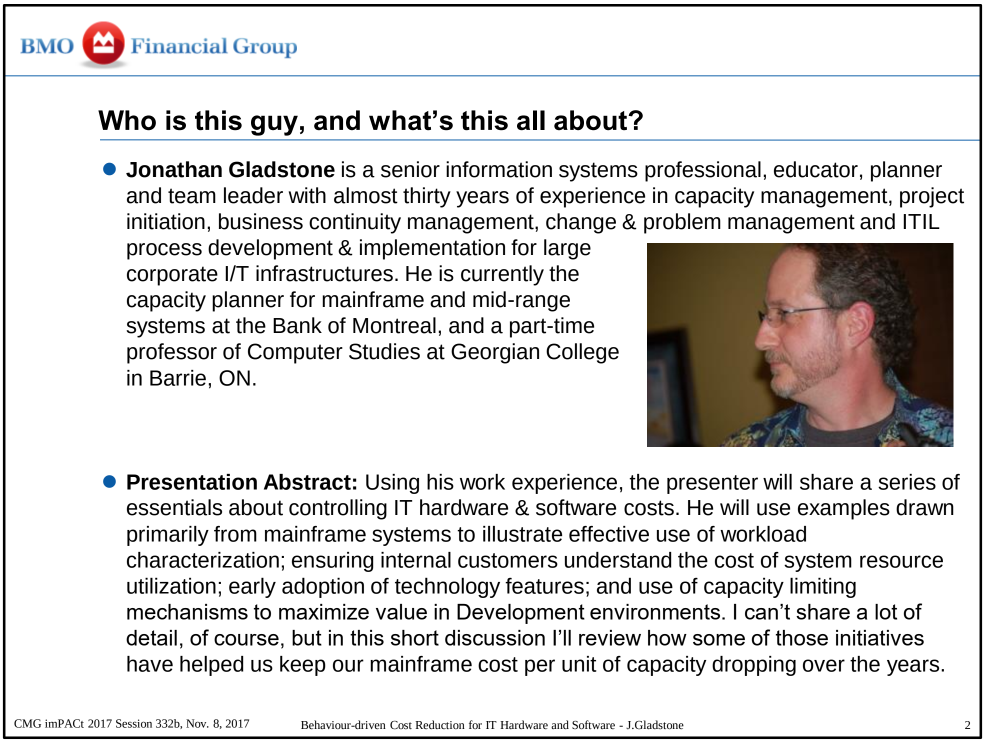


BMO Financial Group

Behaviour-driven Cost Reduction for IT Hardware & Software

Jonathan Gladstone, *P.Eng.*
Capacity Planner, Mainframe & Midrange Systems
Bank of Montreal

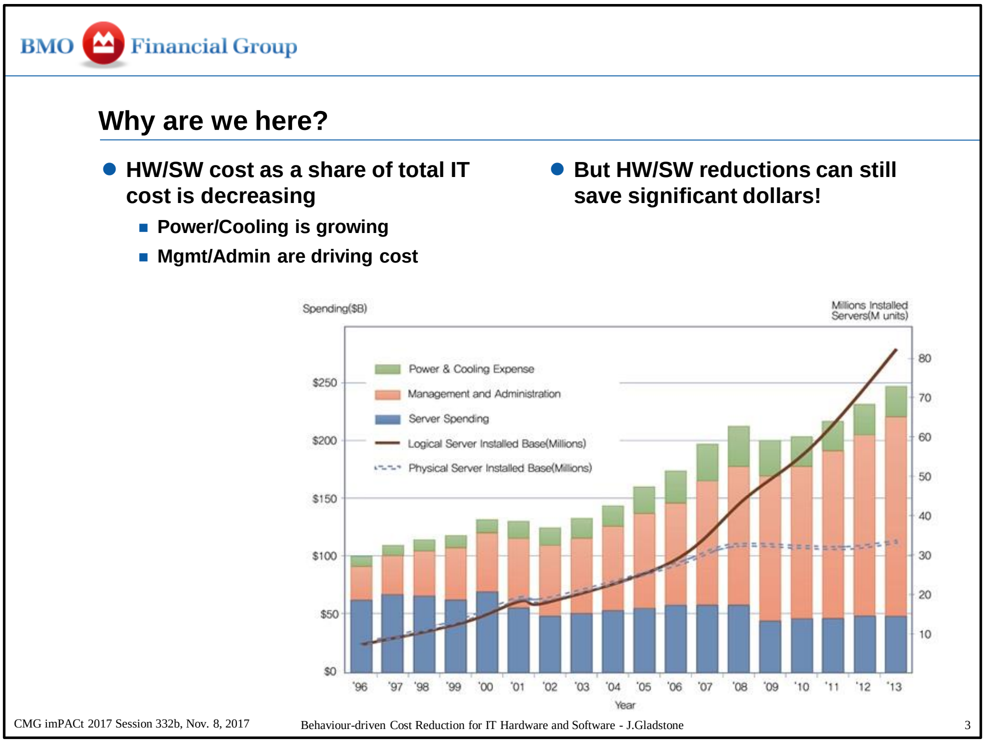
Session 332b
CMG imPACt 2017
Wednesday, November 8 – 11:35-12:05



Who is this guy, and what's this all about?

- **Jonathan Gladstone** is a senior information systems professional, educator, planner and team leader with almost thirty years of experience in capacity management, project initiation, business continuity management, change & problem management and ITIL process development & implementation for large corporate I/T infrastructures. He is currently the capacity planner for mainframe and mid-range systems at the Bank of Montreal, and a part-time professor of Computer Studies at Georgian College in Barrie, ON.

- **Presentation Abstract:** Using his work experience, the presenter will share a series of essentials about controlling IT hardware & software costs. He will use examples drawn primarily from mainframe systems to illustrate effective use of workload characterization; ensuring internal customers understand the cost of system resource utilization; early adoption of technology features; and use of capacity limiting mechanisms to maximize value in Development environments. I can't share a lot of detail, of course, but in this short discussion I'll review how some of those initiatives have helped us keep our mainframe cost per unit of capacity dropping over the years.



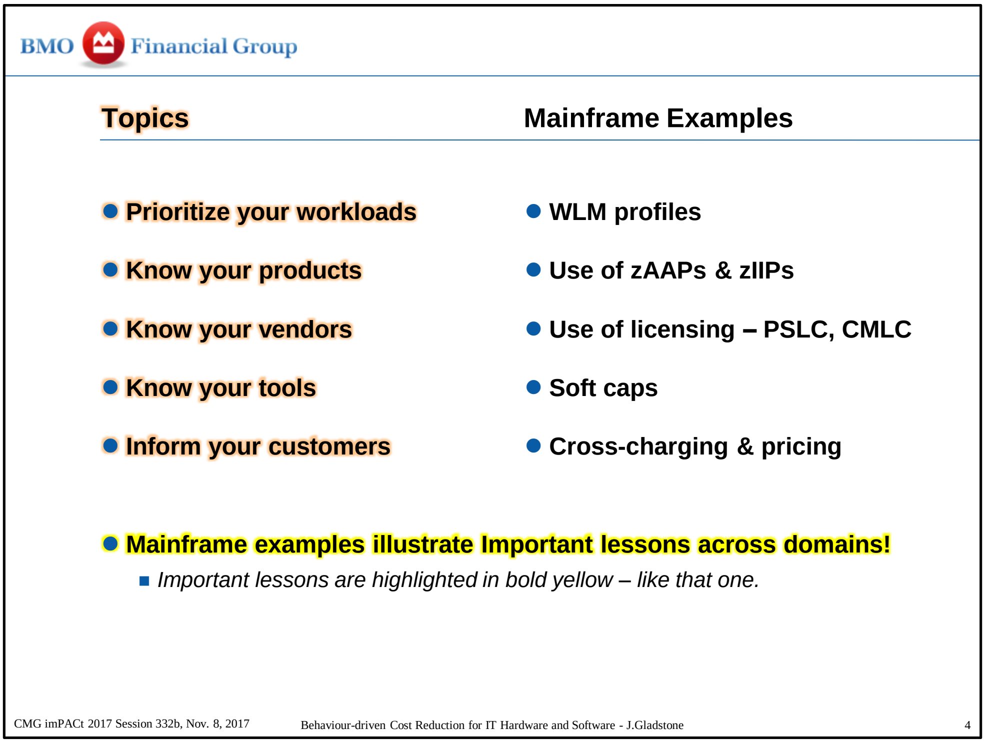
BMO Financial Group

Why are we here?

- HW/SW cost as a share of total IT cost is decreasing
  - Power/Cooling is growing
  - Mgmt/Admin are driving cost
- But HW/SW reductions can still save significant dollars!

Chart from Jang, S.M., An, B.S., On, J.H., Lee, B.G. & Jun, S.I. (August, 2015). “Energy Efficient Computing Technology Trends”, in Electronics & Telecommunications Trends, Vol. 30 No. 4, pp.12-25, August 2015. Retrieved from https://ettrends.etri.re.kr/ettrends/pubreader.do?volume=30&issue=4&page=12&paperno=0905002051 on May 23, 2017.

We can show our value in Performance Management & Capacity Planning by reducing cost drivers anywhere
This chart is about “servers”, which probably means distributed systems (Windows, Linux & Unix). We can only make direct reductions against 20% of overall cost with HW/SW actions… but the other two categories are driven partly by numbers of images and devices so there’s an indirect effect too. And some of the improvements involve automation, so go directly against “Big Orange”, which is primarily staff cost.



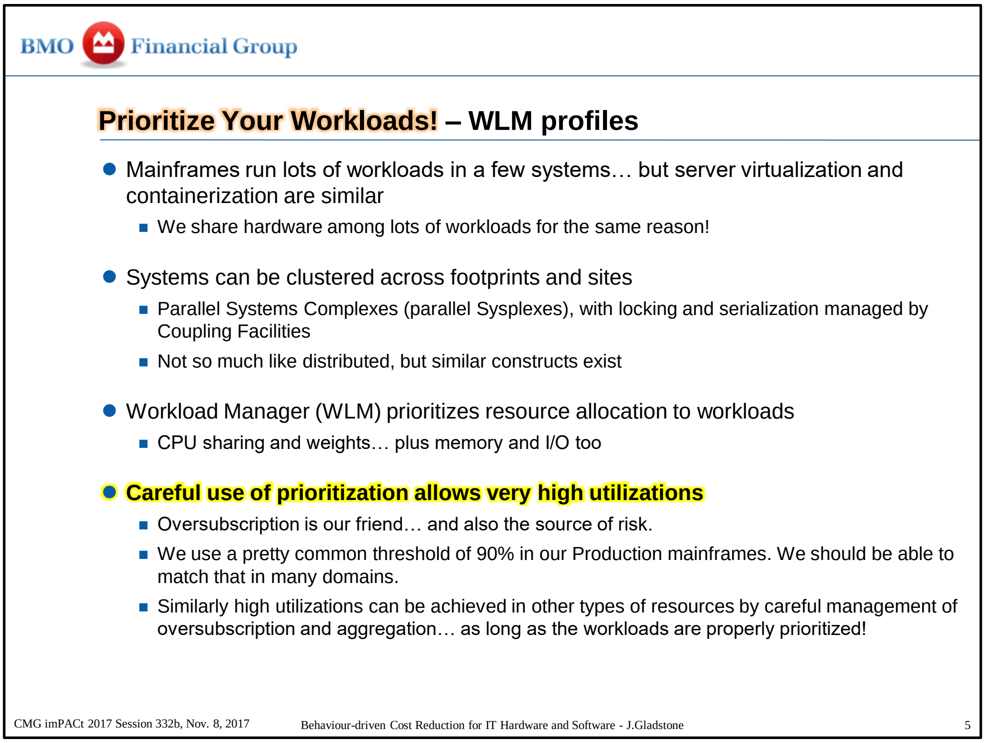
BMO Financial Group

Topics	Mainframe Examples
● **Prioritize your workloads**	● **WLM profiles**
● **Know your products**	● **Use of zAAPs & zIIPs**
● **Know your vendors**	● **Use of licensing – PSLC, CMLC**
● **Know your tools**	● **Soft caps**
● **Inform your customers**	● **Cross-charging & pricing**

- **Mainframe examples illustrate Important lessons across domains!**
  - *Important lessons are highlighted in bold yellow – like that one.*

Why all the mainframe examples? Two significant reasons:

1) The guy giving the presentation has spent the largest chunk of his career – over 20 years out of just under 30 – working with mainframes.
2) Mainframes have by far the most mature Capacity Management processes. We've learned the lessons here over and over again. One of the problems in our industry – as in many others- is a failure to apply lessons learned in one area (whether that's a domain, like distributed versus mainframe servers, or an industry, like aircraft piloting versus medicine) to another. Let's learn from one another! If you're not convinced, here's a pitch: read "The Checklist Manifesto" by Dr. Atul Gawande (Metropolitan Books, Henry Holt & Co., NY, NY; 2009. ISBN 0805091742). A colleague of ours, David van Geilswyk (an independent consultant formerly with Mid-Range Computer Group and Blair Technology in the Toronto area) recommended it to me; it's a lesson in learning lessons!



BMO Financial Group

Prioritize Your Workloads! – WLM profiles

- Mainframes run lots of workloads in a few systems… but server virtualization and containerization are similar
  - We share hardware among lots of workloads for the same reason!
- Systems can be clustered across footprints and sites
  - Parallel Systems Complexes (parallel Sysplexes), with locking and serialization managed by Coupling Facilities
  - Not so much like distributed, but similar constructs exist
- Workload Manager (WLM) prioritizes resource allocation to workloads
  - CPU sharing and weights… plus memory and I/O too
- **Careful use of prioritization allows very high utilizations**
  - Oversubscription is our friend… and also the source of risk.
  - We use a pretty common threshold of 90% in our Production mainframes. We should be able to match that in many domains.
  - Similarly high utilizations can be achieved in other types of resources by careful management of oversubscription and aggregation… as long as the workloads are properly prioritized!

CMG imPACt 2017 Session 332b, Nov. 8, 2017 Behaviour-driven Cost Reduction for IT Hardware and Software - J.Gladstone 5

In mainframes, dozens or hundreds of heterogeneous workloads (thousands of tasks) run together in a single system image. System images run in logical partitions (LPARs), which are containers of virtual resources.

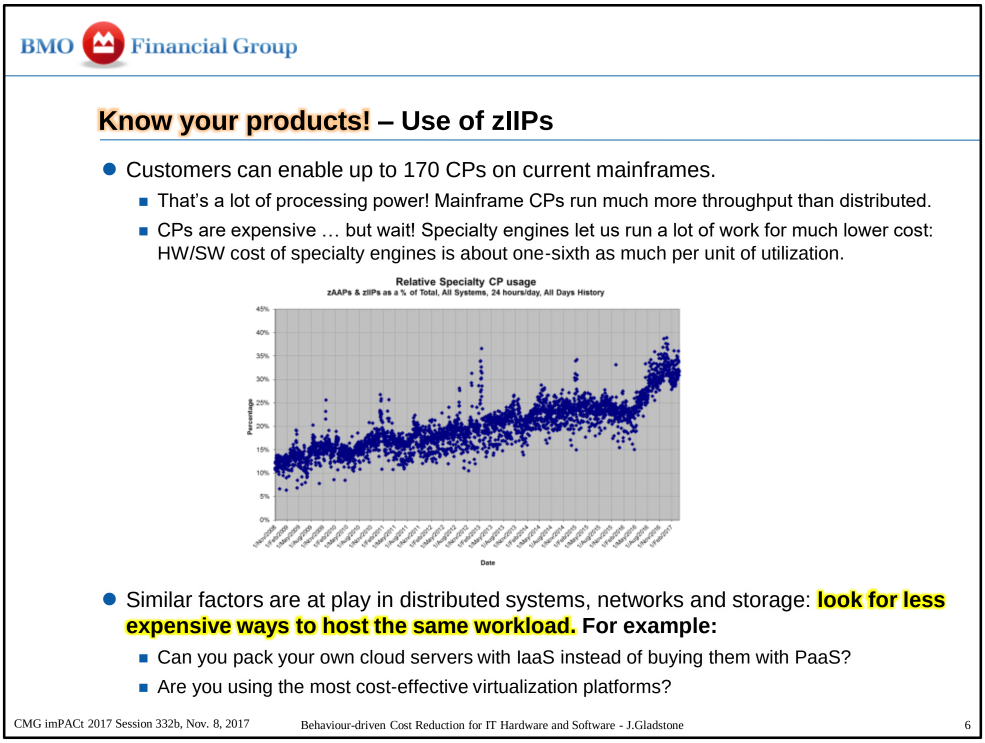
- Yes, just like distributed! Or really, vice versa – mainframes have done this since the 1970s.
- Even with large numbers of heterogeneous workloads, virtualization allows sharing of resources while still providing ways to prevent data spillage.

Groups of system images are clustered together across disparate physical footprints (sometimes at multiple sites) into Parallel Systems Complexes (parallel Sysplexes), with locking and serialization managed by Coupling Facilities

- Clustering also exists in distributed systems, but works differently.

Workload Manager (WLM) prioritizes the many heterogeneous workloads across all of the systems in a given Sysplex with various classes and targets for batch and online workloads

- Processor Resource / System Manager (PR/SM) is a type 1 hypervisor that works across LPARs on a footprint. WLM manages workloads across the Sysplex cluster. Intelligent Resource Director (IRD) can dynamically manage resource allocation to optimize workload throughput based on WLM targets.
- WLM can manage memory and I/O access too, using the same prioritization.
- These function similarly to distributed systems hypervisors, and in some respects also to prioritization systems for networks, storage and (perhaps to a lesser extent) converged/hyperconverged systems.



Know your products! – Use of zIIPs

- Customers can enable up to 170 CPs on current mainframes.
  - That's a lot of processing power! Mainframe CPs run much more throughput than distributed.
  - CPs are expensive … but wait! Specialty engines let us run a lot of work for much lower cost: HW/SW cost of specialty engines is about one-sixth as much per unit of utilization.

Relative Specialty CP usage
zAAPs & zIIPs as a % of Total, All Systems, 24 hours/day, All Days History

- Similar factors are at play in distributed systems, networks and storage: **look for less expensive ways to host the same workload. For example:**
  - Can you pack your own cloud servers with IaaS instead of buying them with PaaS?
  - Are you using the most cost-effective virtualization platforms?

CMG imPACt 2017 Session 332b, Nov. 8, 2017 Behaviour-driven Cost Reduction for IT Hardware and Software - J.Gladstone 6

Customers can enable up to 170 engines on a current generation z14 footprint.

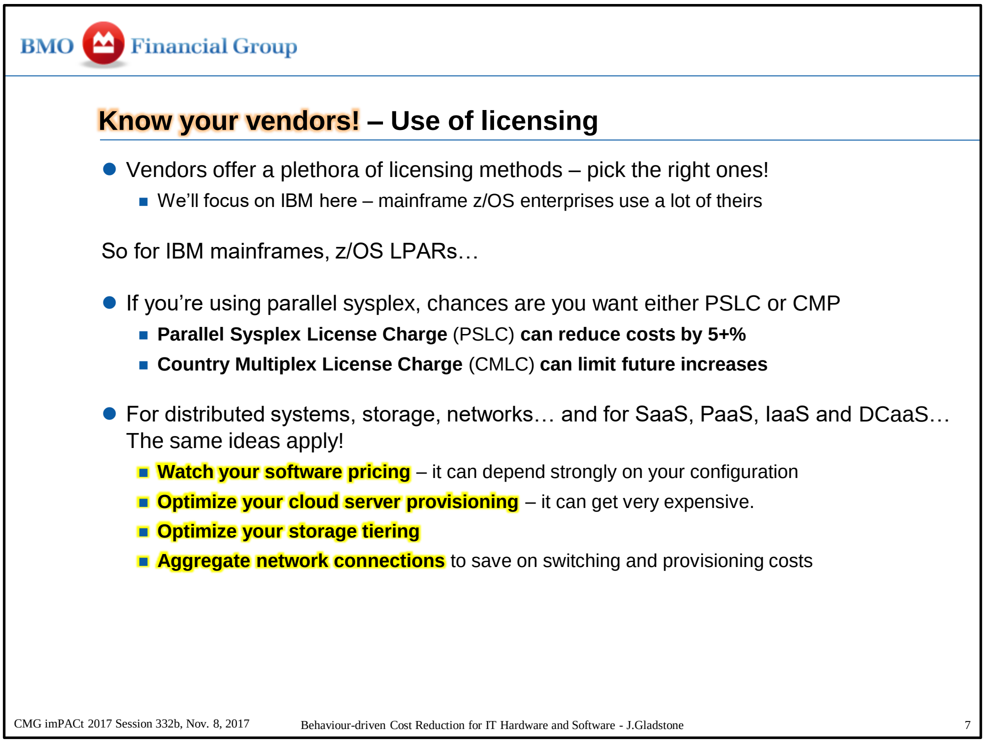
- That's a lot of processing power! For comparative purposes, one engine can run anywhere from four or five to fifteen or twenty production Linux images, depending on workload.

z Systems Integrated Information Processors (zIIPs) are just the same engines with a different label, but they can run a significant subset of the work for much lower cost.

- General-purpose CP hardware costs well over C$500k for hardware, and usually a lot more again for software. zIIP hardware costs less than C$100k, with no software cost. Add shared infrastructure costs (box, memory, crypto, I/O…), cost ratio is about one-sixth per unit of utilization. **For cost optimization, we continue to work to maximize our zIIP utilization.**
- In distributed systems, some workloads will run better in Linux than Windows, or vice versa. Some processor designs may allow off-loading some work to less-expensive processors. These and similar factors may allow significant savings.

Cost optimization means choosing your platforms and directing your workloads to lower cost while maintaining availability. It's the same balancing act as always.

- When you're purchasing cloud processing power, you can significantly reduce cost by buying 'bare metal' (IaaS) and installing your own hypervisor to take advantage of oversubscription. If you buy PaaS instead, the vendor gets that advantage (see also next page).
- For Linux workloads, IBM Z mainframe hardware is an option. Integrated Facility for Linux (IFL) CPs are again just the same engines. The hardware costs are low, like zIIPs, but there are also software and hypervisor costs. Even so… one way to reduce your Linux costs is to consider running RHEL (Red Hat), SLES (SUSE), Ubuntu or Canonical natively on IFLs or under z/VM or KVM hypervisors.
- For storage, tiered solutions with lots of cache can reduce cost while increasing (apparent) speed – and improve service and save on processing cost too by reducing wait times.



BMO Financial Group

Know your vendors! – Use of licensing

- Vendors offer a plethora of licensing methods – pick the right ones!
  - We'll focus on IBM here – mainframe z/OS enterprises use a lot of theirs

So for IBM mainframes, z/OS LPARs…

- If you're using parallel sysplex, chances are you want either PSLC or CMP
  - **Parallel Sysplex License Charge** (PSLC) **can reduce costs by 5+%**
  - **Country Multiplex License Charge** (CMLC) **can limit future increases**
- For distributed systems, storage, networks… and for SaaS, PaaS, IaaS and DCaaS… The same ideas apply!
  - **Watch your software pricing** – it can depend strongly on your configuration
  - **Optimize your cloud server provisioning** – it can get very expensive.
  - **Optimize your storage tiering**
  - **Aggregate network connections** to save on switching and provisioning costs

CMG imPACt 2017 Session 332b, Nov. 8, 2017 Behaviour-driven Cost Reduction for IT Hardware and Software - J.Gladstone 7

Each of us uses a variety of vendors, for hardware, software and services…

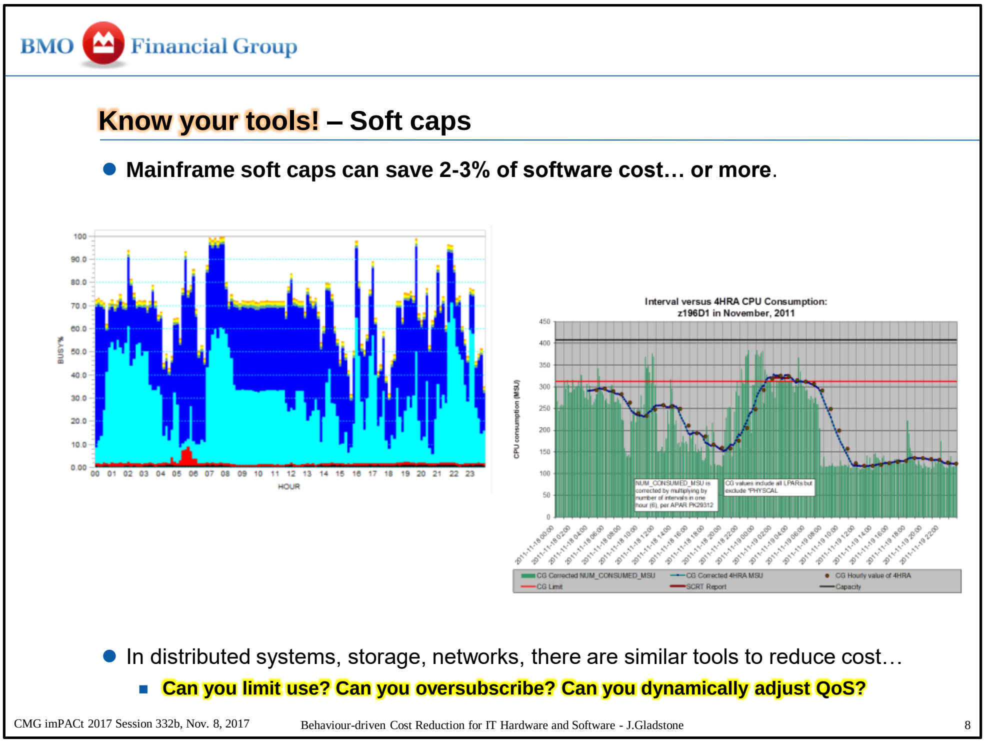
If you're using parallel sysplex, chances are you want either PSLC or CMLC

- **Parallel Sysplex License Charge** (PSLC) **can reduce costs by 5+%** – for select products (including z/OS, DB2, CICS, IMS and many others) allows customers to pay a single software charge per product based on monthly utilization, and for a single sysplex per box as long as that sysplex has a system using over 50% of each given box; pay based on peak four-hour rolling average (4HRA) incurred each month. Priced per footprint.
- **Country Multiplex License Charge** (CMLC) **can limit future increases** – pay for combined usage across footprints anywhere within a given country. Price point based on existing customer usage for recent three-month period, but later on buffers against workload variability

Optimization:

- Software pricing can vary a lot depending on the size and number of images or CPs or physical servers implemented. You can often achieve considerable savings by changing your configuration to minimize the cost drivers for specific kinds of software. Oracle and SAS are good examples here.
- Optimizing cloud provisioning… all of these can go in either direction, just add "vice versa" to the end of each.
  - If you're using SaaS and lots of customisation services, should you consider moving to PaaS and managing your own software?
  - If you're using PaaS for lots of server images, should you consider moving to IaaS and managing from the hypervisor level up, so you can share resources more effectively?
  - If you're using IaaS for lots of capacity (storage and/or compute), should you consider moving to DCaaS and managing from the hardware level up?
- Potential cost reductions from optimizing storage tiering are self-evident. This applies to both on-premises and external cloud enterprise storage offerings.
- Potential cost reductions from network aggregation are self-evident too.

In all cases, there's a management cost for achieving savings. For aggregation or virtualization, there's also the risk involved in oversubscription and the cost of mitigating that risk.



BMO Financial Group

Know your tools! – Soft caps

- **Mainframe soft caps can save 2-3% of software cost… or more**.

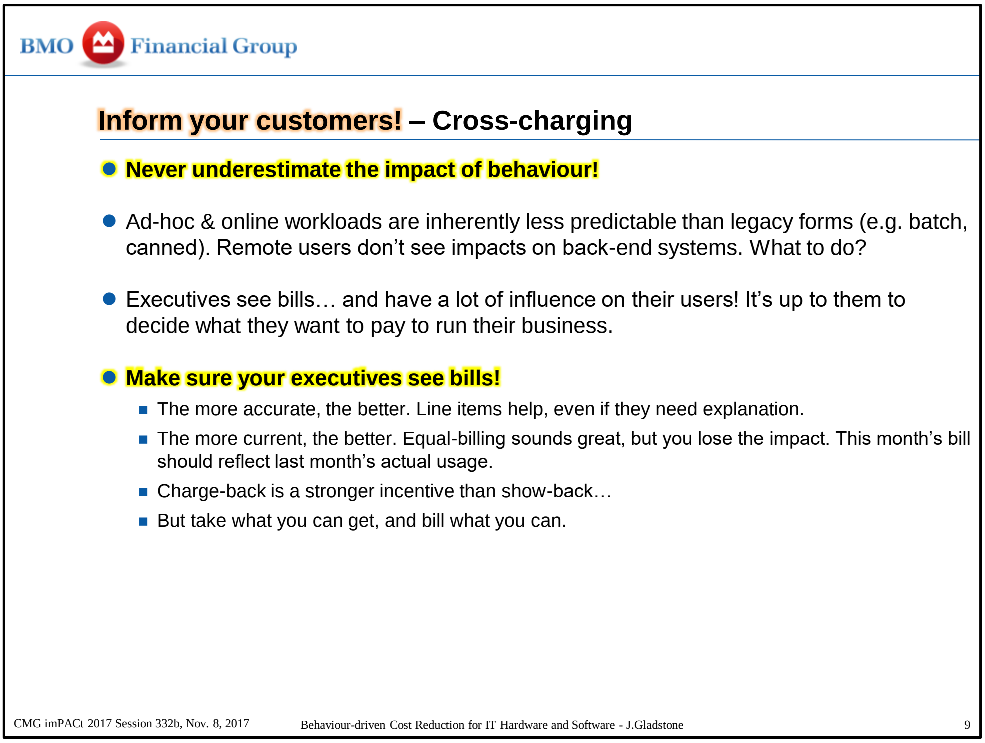
- In distributed systems, storage, networks, there are similar tools to reduce cost…
  - **Can you limit use? Can you oversubscribe? Can you dynamically adjust QoS?**

CMG imPACt 2017 Session 332b, Nov. 8, 2017 Behaviour-driven Cost Reduction for IT Hardware and Software - J.Gladstone 8

IBM allows customers to limit utilization of groups of one or more LPARs within each footprint to less than their instantaneous capacity.

- Capping is based on the same four-hour rolling averages used by PSLC and CMLC.
- Capping an entire box limits the software charge to the cap level regardless of instantaneous or actual 4HRA utilization, for most IBM mainframe software and some other vendors too.

Limiting use or oversubscribing or adjusting QoS can be set at whatever level you think is right. More risk and more management & mitigation gets you lower cost. The more pain, the more gain!



BMO Financial Group

Inform your customers! – Cross-charging

- **Never underestimate the impact of behaviour!**
- Ad-hoc & online workloads are inherently less predictable than legacy forms (e.g. batch, canned). Remote users don’t see impacts on back-end systems. What to do?
- Executives see bills… and have a lot of influence on their users! It’s up to them to decide what they want to pay to run their business.
- **Make sure your executives see bills!**
  - The more accurate, the better. Line items help, even if they need explanation.
  - The more current, the better. Equal-billing sounds great, but you lose the impact. This month’s bill should reflect last month’s actual usage.
  - Charge-back is a stronger incentive than show-back…
  - But take what you can get, and bill what you can.

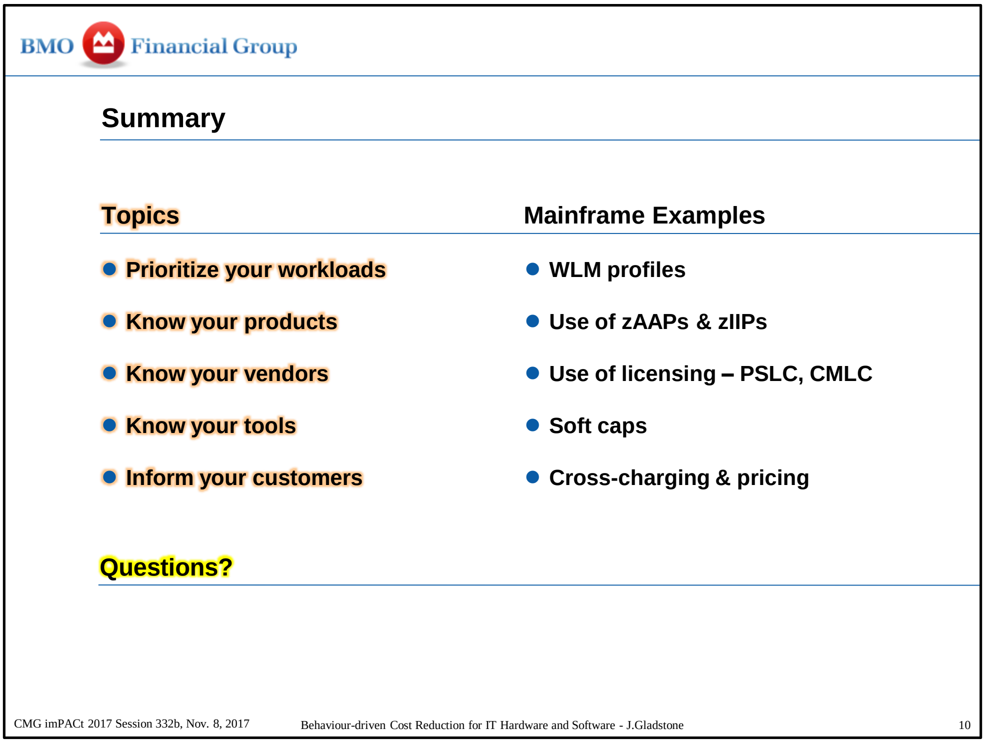
CMG imPACt 2017 Session 332b, Nov. 8, 2017 Behaviour-driven Cost Reduction for IT Hardware and Software - J.Gladstone 9

It costs time and money to build and maintain an accurate, effective charge-back (or show-back) system. Why invest?
CONS

- Time and money spent on systems, agents, data collection & storage, report creation etc.
- Difficulty of establishing units of charge, rates, SLAs and consequences for missing SLAs.
- Arguments from bill recipients.
- Dissatisfaction with rates etc.

PROS

- Arguments from bill recipients. Yes, this is present in both Pros and Cons! This is your chance to explain what drives bills… so your biggest chance to influence utilization behaviours.
- Establishes accountability for costs with customer-facing lines of business – IT is now an internal service provider, instead of ‘just a cost centre’.
- Paying for bills drives better behaviour. For an analogy, think of water and power use in a building where utilities are included with rent or fees, compared with a building where each resident pays for their own utility use.



BMO Financial Group

Summary

Topics	Mainframe Examples
Prioritize your workloads	WLM profiles
Know your products	Use of zAAPs & zIIPs
Know your vendors	Use of licensing – PSLC, CMLC
Know your tools	Soft caps
Inform your customers	Cross-charging & pricing

Questions?