

Exploring IT Cost Components – How to Maximize your IT Investments

Ray Jones

Vice President, z System Worldwide Software

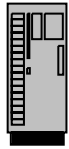


Many Cost Components

80:20 rule helps to achieve reasonable results in a short time

Components

Hardware



List vs Discounted

Fully configured vs. basic, Prod. vs. DR

Refresh / upgrade, Solution Edition...

Software



IBM and ISV, OTC and Annual maint (S&S)

MLC, PVU, RVU, ELA, core, system

People



FTE rate, in house vs. contract

Network



Adapters, switches, routers, hubs

Charges, Allocated or apportioned, understood or clueless

Storage



ECKD, FBA, SAN, Compressed, Primary, secondary

Disk (multiple vendors), tape, Virtual, SSD

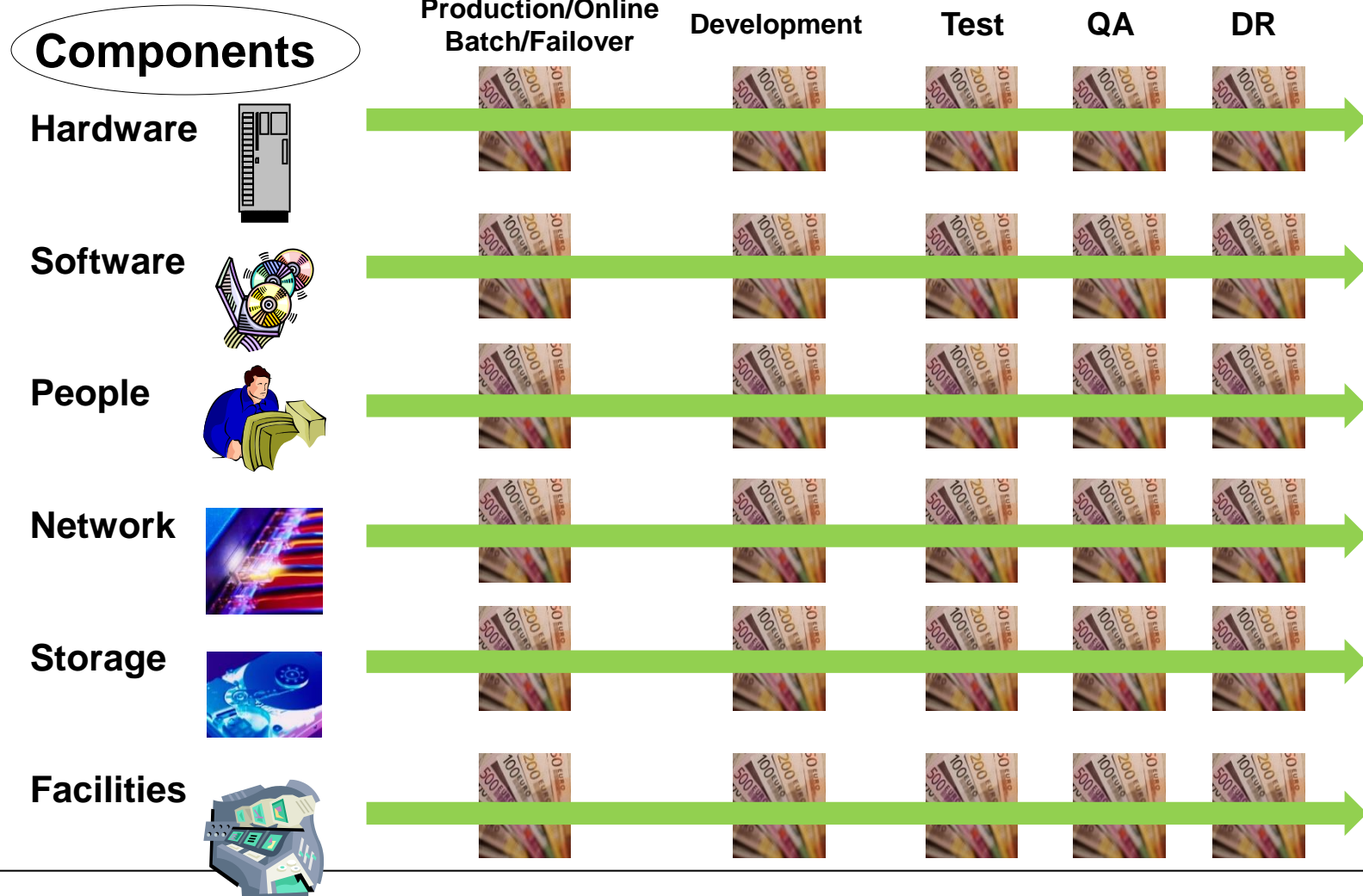
Facilities



Space, electricity, air cooling, infrastructure including UPS and generators, alternate site(s), bandwidth

Environments Multiply Components

Environments



Time Factors Drive Growth And Cost

- Migration time and effort
- Business organic growth and/or planned business changes affect capacity requirements
 - e.g. Change of access channel or adding a new internet accessible feature can double or triple a components workload
 - Link a business metric (e.g. active customer accounts) to workload (e.g. daily transactions) and then use business inputs to drive the TCO case
- Other periodic changes – hardware refresh or software remediation

Non-Functional Requirements Can Drive Additional Resource Requirements

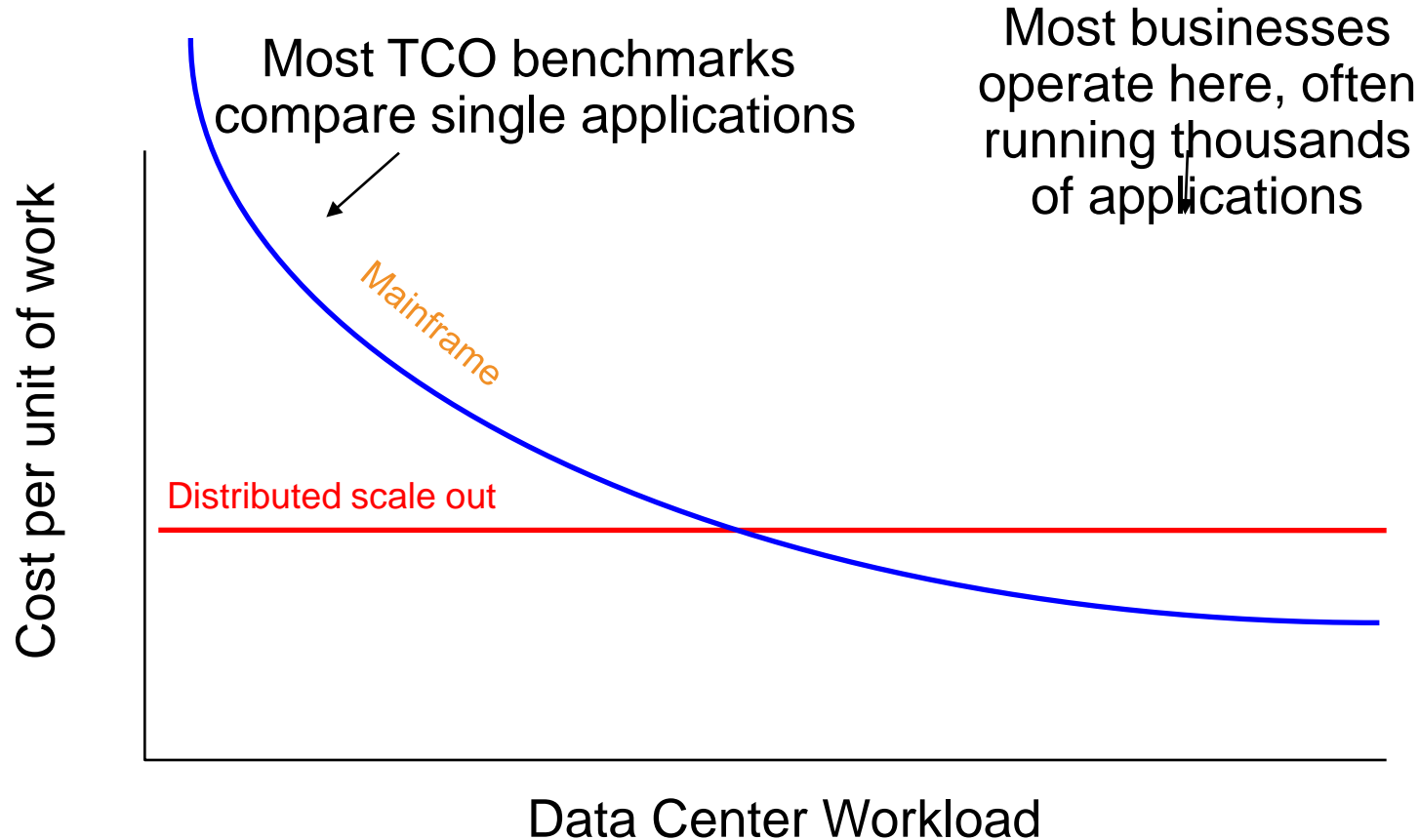


Availability ... Security ... Resiliency ... Scalability ...

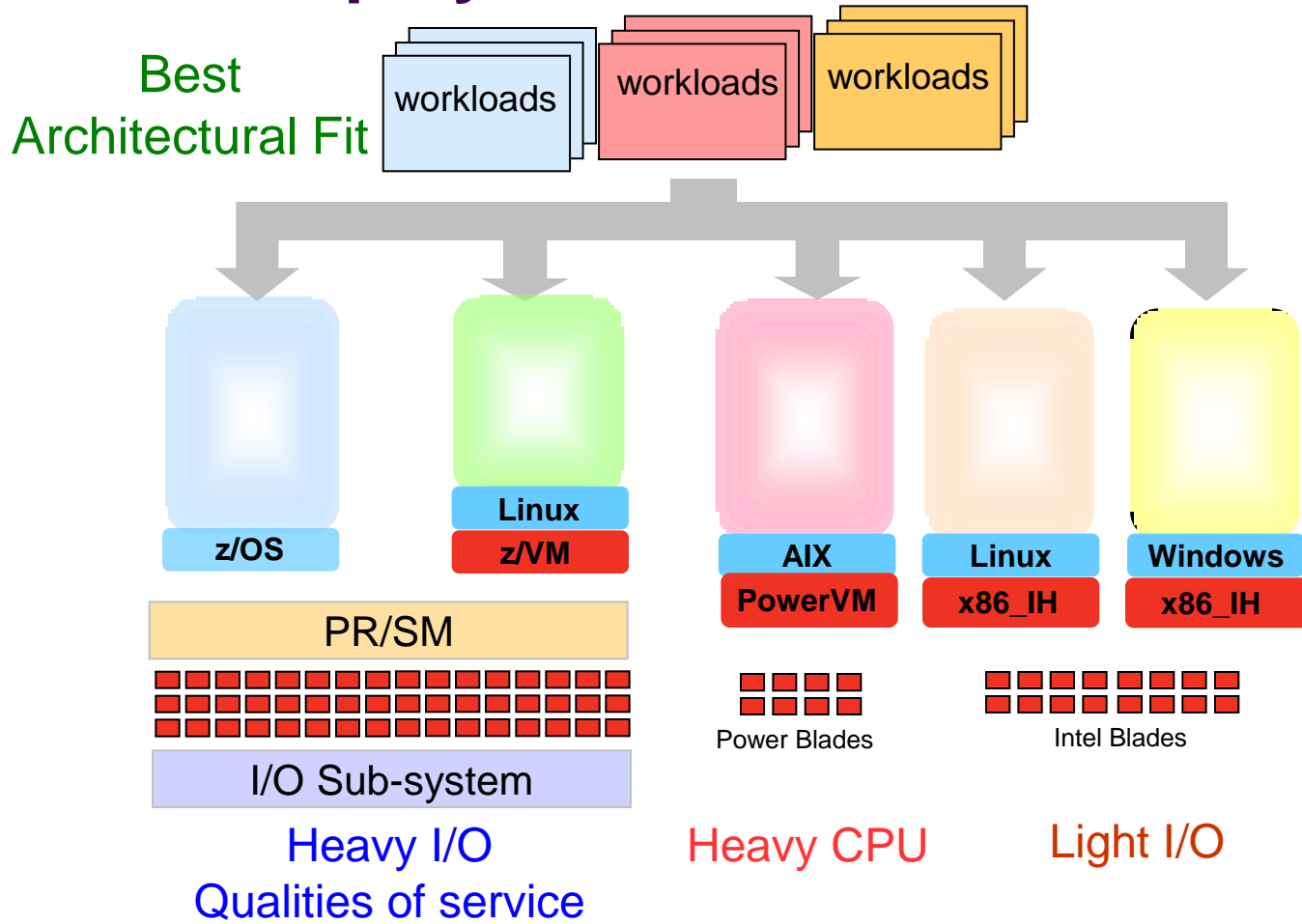


Qualities of Service, Non-Functional Requirements

Mainframe Cost/Unit of Work Decreases as Workload Increases



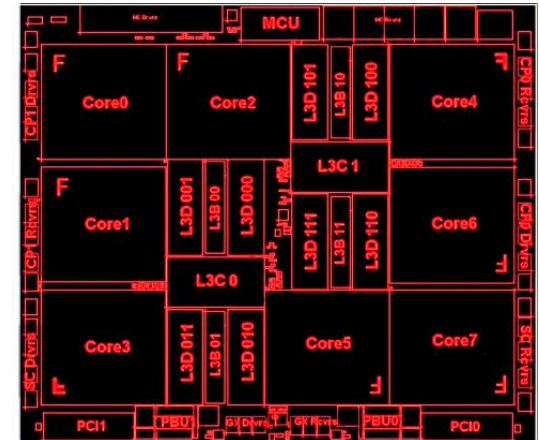
Workload Characteristics Influence The Best Fit Deployment Decision



Deploy or consolidate workloads on the environment best suited for each workload to yield lowest cost

Designed for transaction processing and data serving

- New **8-core** Processor Design in **22nm Silicon Technology**
- **Optimized Instruction Processing** (Out-of-Order Execution Pipeline, Relative-branch execution units, Software Prefetch Directives)
- Larger **caches to optimize** data serving environments
- **Architecture Extensions** (Transactional Execution, RDMA, Runtime Instrumentation) provide enhanced workload performance
- Substantial economies of scale with **simultaneous multi-threading delivering more throughput** for Linux and zIIP-eligible workloads
- **Single Instruction Multiple Data (SIMD)** improves performance of complex mathematical models
- Up to 2X **improved cryptographic performance** with enhanced Central Processor Assist for Cryptographic Functions (CPACF)
- **Compress more data** helping to save disk space and cut data transfer time with improved **on chip hardware compression**

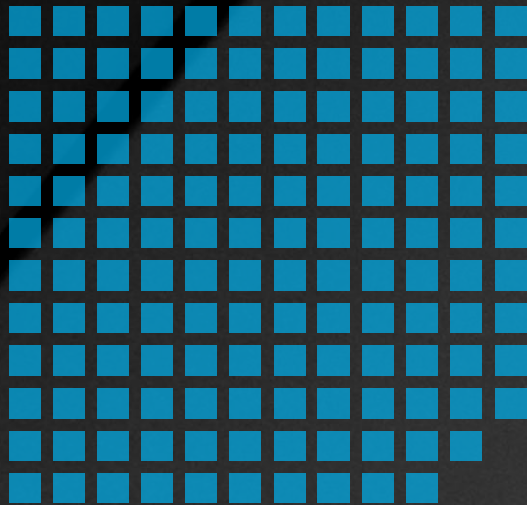


Plus 10 TB of memory to further improve performance

Balanced System Design

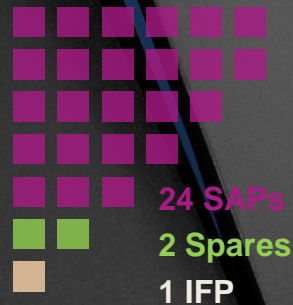
I/O and coprocessors bring added compute power to workloads

Up to 141 cores
on a CPC

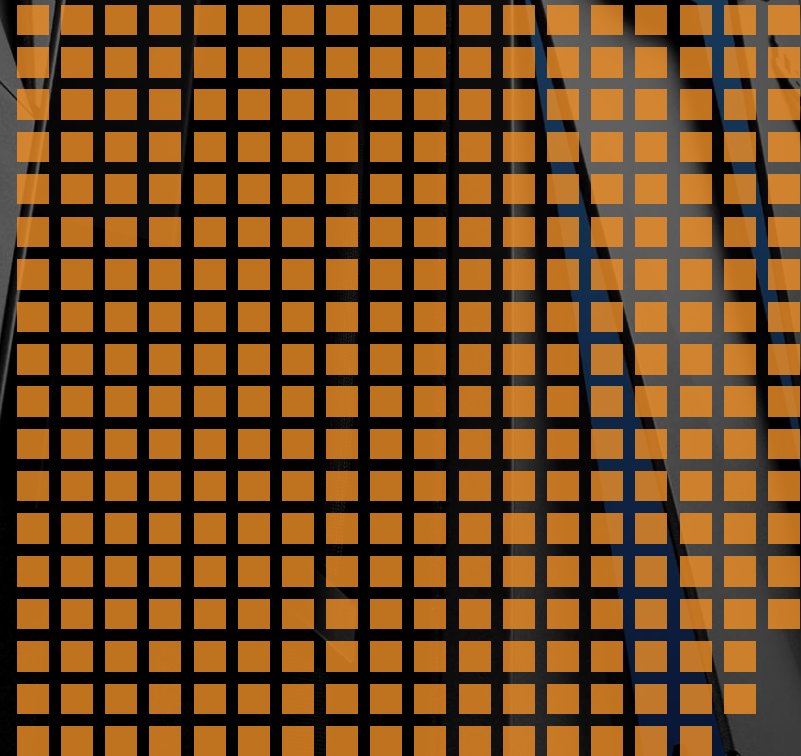


- Share up to 141 processors with up to 85 LPARS
- Configure the processors as CPs, IFLs, zIIPs, or ICFs

Up to 27 cores for
offload system
processing



Plus up to 640 POWER cores:
I/O and Coprocessors



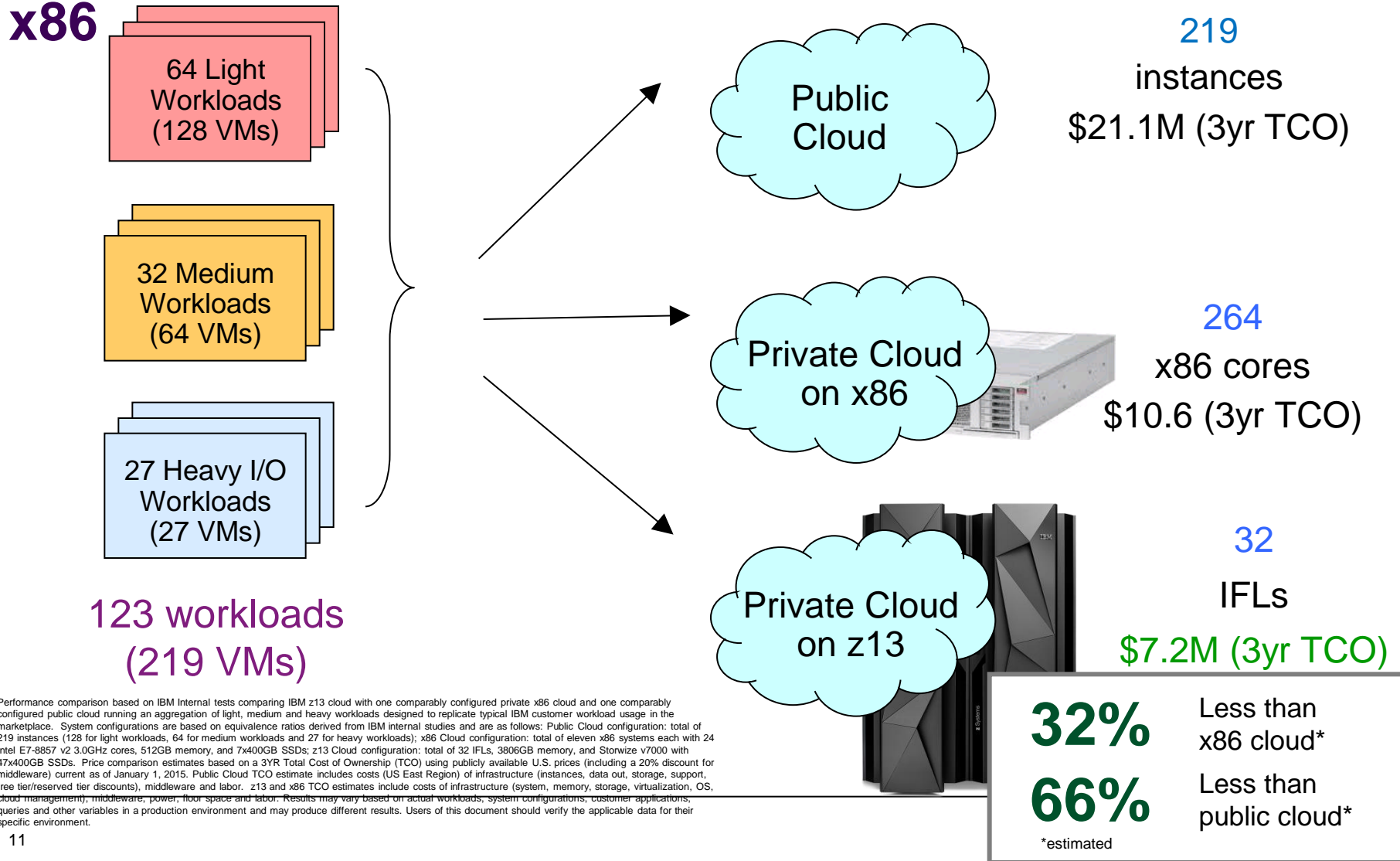
Resilient and intelligent I/O

- New FICON Express16S links reduce latency for workloads such as DB2
- **Reduce up to 43% of DB2 write operations with IBM zHyperWrite** – technology for DS8000 and z/OS for Metro Mirror environment
- First system to use a **standards based approach for enabling Forward Error Correction** for a complete end to end solution
- Clients with multi-site configurations can expect **I/O service time improvement** when writing data remotely which can **benefit GDPS or TPC-R HyperSwap**
- **Extend z/OS workload management policies into SAN fabric** to manage the network congestion
- New Easy Tier API removes requirement from application/administrator to manage hardware resources



Optimized for enterprise-scale data from multiple platforms and devices

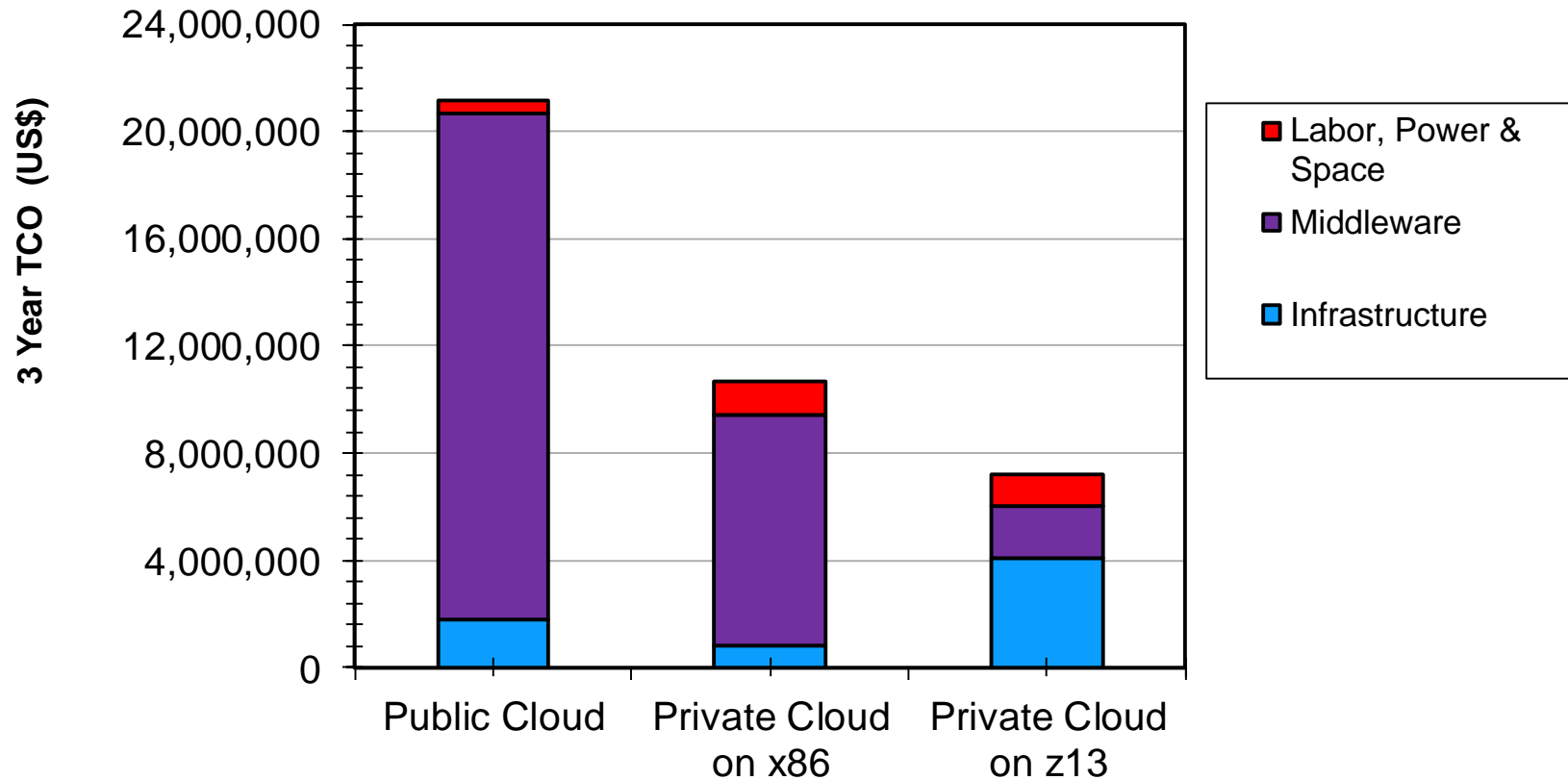
A private cloud on z13 yields the lowest TCO compared to a public cloud and a private cloud on x86



Performance comparison based on IBM Internal tests comparing IBM z13 cloud with one comparably configured private x86 cloud and one comparably configured public cloud running an aggregation of light, medium and heavy workloads designed to replicate typical IBM customer workload usage in the marketplace. System configurations are based on equivalence ratios derived from IBM internal studies and are as follows: Public Cloud configuration: total of 219 instances (128 for light workloads, 64 for medium workloads and 27 for heavy workloads); x86 Cloud configuration: total of eleven x86 systems each with 24 Intel E7-8857 v2 3.0GHz cores, 512GB memory, and 7x400GB SSDs; z13 Cloud configuration: total of 32 IFLs, 3806GB memory, and Storwize v7000 with 47x400GB SSDs. Price comparison estimates based on a 3YR Total Cost of Ownership (TCO) using publicly available U.S. prices (including a 20% discount for middleware) current as of January 1, 2015. Public Cloud TCO estimate includes costs (US East Region) of infrastructure (instances, data out, storage, support, free tier/reserved tier discounts), middleware and labor. z13 and x86 TCO estimates include costs of infrastructure (system, memory, storage, virtualization, OS, cloud management), middleware, power, floor space and labor. Results may vary based on actual workloads, system configurations, customer applications, queries and other variables in a production environment and may produce different results. Users of this document should verify the applicable data for their specific environment.

A breakdown shows how middleware costs soar on both the x86 cloud and the public cloud

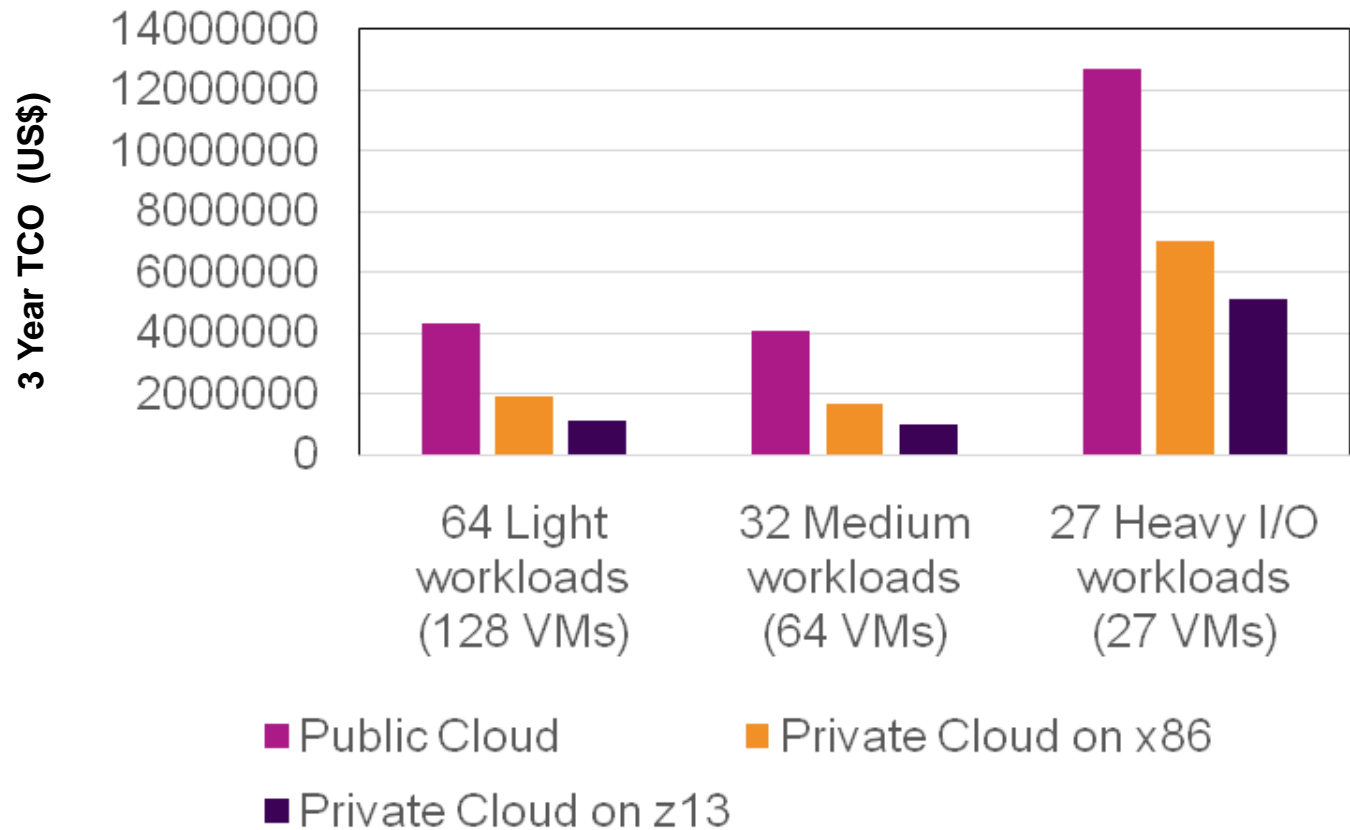
Case Study: 123 Workloads (219 VMs)



Performance comparison based on IBM Internal tests comparing IBM z13 cloud with one comparably configured private x86 cloud and one comparably configured public cloud running an aggregation of light, medium and heavy workloads designed to replicate typical IBM customer workload usage in the marketplace. System configurations are based on equivalence ratios derived from IBM internal studies and are as follows: Public Cloud configuration: total of 219 instances (128 for light workloads, 64 for medium workloads and 27 for heavy workloads); x86 Cloud configuration: total of eleven x86 systems each with 24 Intel E7-8857 v2 3.0GHz cores, 512GB memory, and 7x400GB SSDs; z13 Cloud configuration: total of 32 IFLs, 3806GB memory, and Storwize v7000 with 47x400GB SSDs. Price comparison estimates based on a 3YR Total Cost of Ownership (TCO) using publicly available U.S. prices (including a 20% discount for middleware cores) current as of January 1, 2015. Public Cloud TCO estimate includes costs (US East Region) of infrastructure (instances, data out, storage, support, free tier/reserved tier discounts), middleware and labor. z13 and x86 TCO estimates include costs of infrastructure (system, memory, storage, virtualization, OS, cloud management), middleware, power, floor space and labor. Results may vary based on actual workloads, system configurations, customer applications, queries and other variables in a production environment and may produce different results. Users of this document should verify the applicable data for their specific environment.

A private cloud on z13 yields lowest TCO for a variety of workloads

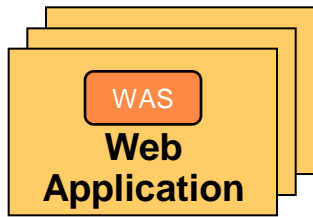
TCO comparison of three types of workloads



Example: Compute intensive non-critical web workloads

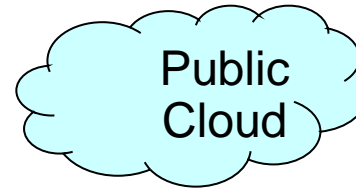
24 Web Workloads

24 VMs total

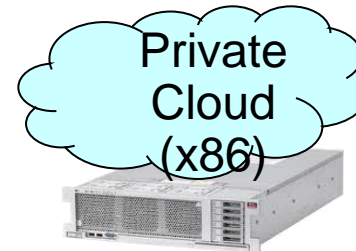


Compute-intensive web workloads

1 VMs per workload;
Each VM requiring 4GB memory; 20GB storage



24 instances
(with total 48 vCPU)



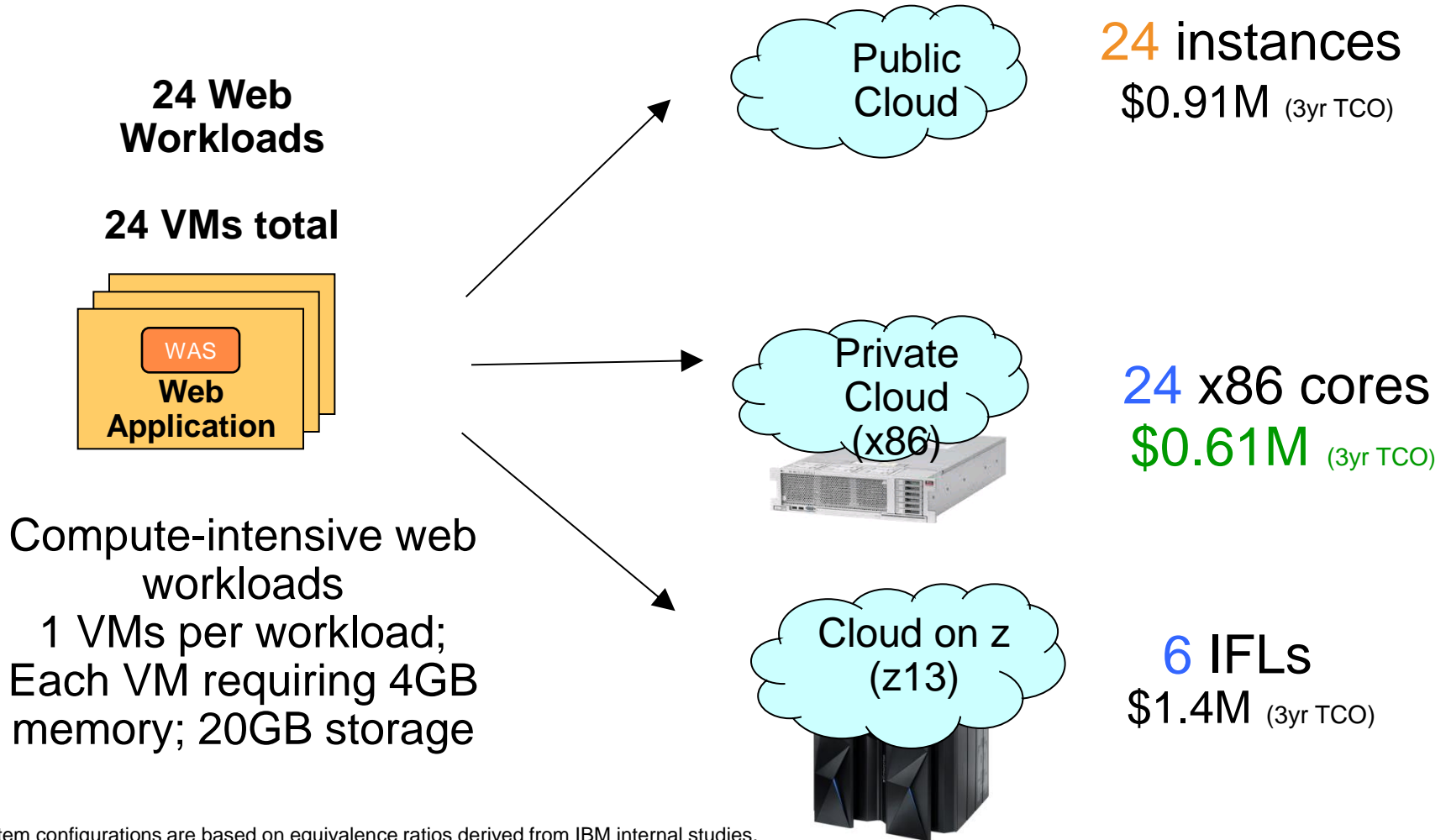
24
x86 cores



6
IFLs

System configurations are based on equivalence ratios derived from IBM internal studies.
Average utilization of 24-core x86 system is assumed to be 50%; avg utilization of z13 with 6 IFLs is assumed to be 75%; transaction response time is the same on all platforms

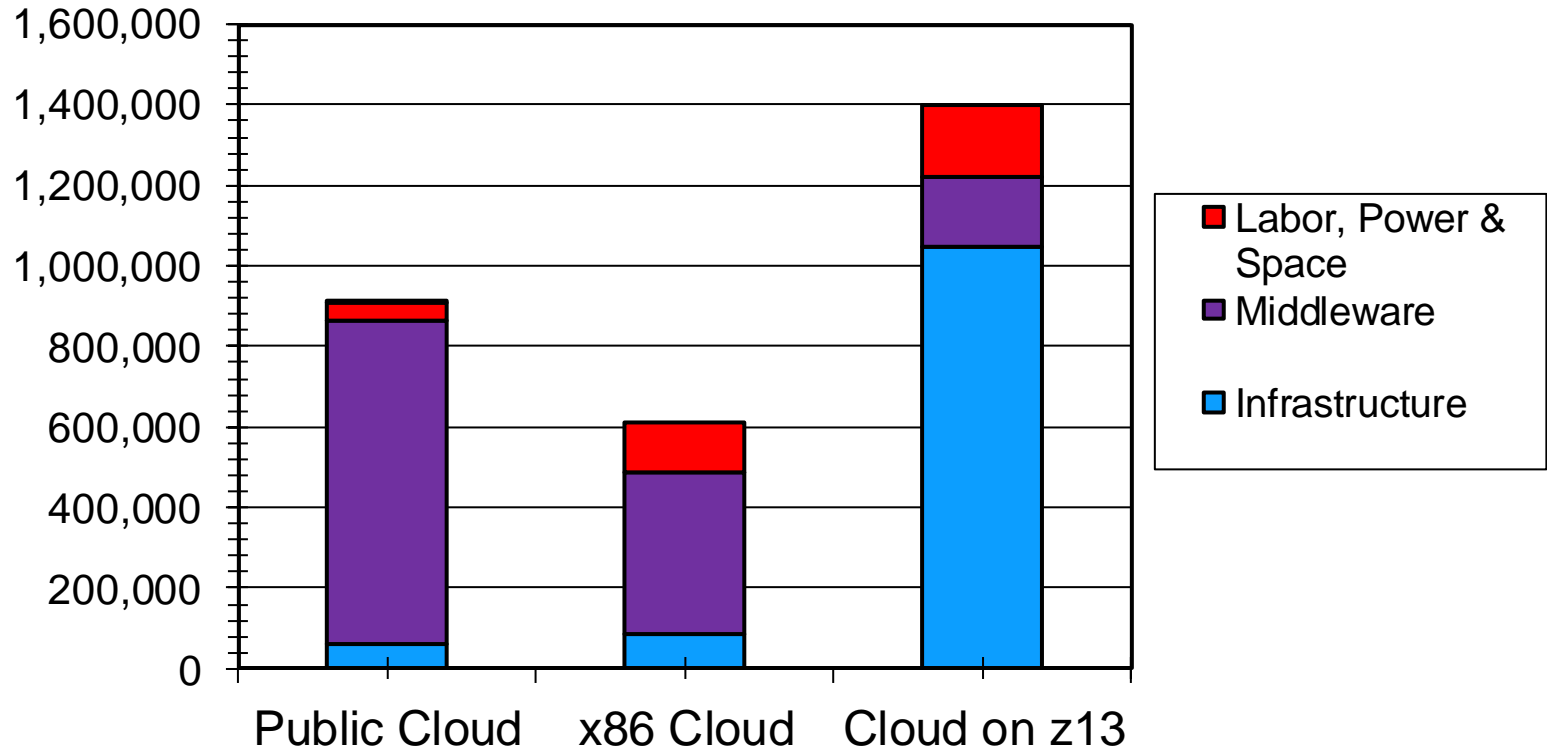
x86 and public cloud yield lower 3yr TCO



System configurations are based on equivalence ratios derived from IBM internal studies. Average utilization of 24-core x86 system is assumed to be 60%; avg utilization of z13 with 6 IFLs is assumed to be 75%; transaction response time is the same on all platforms

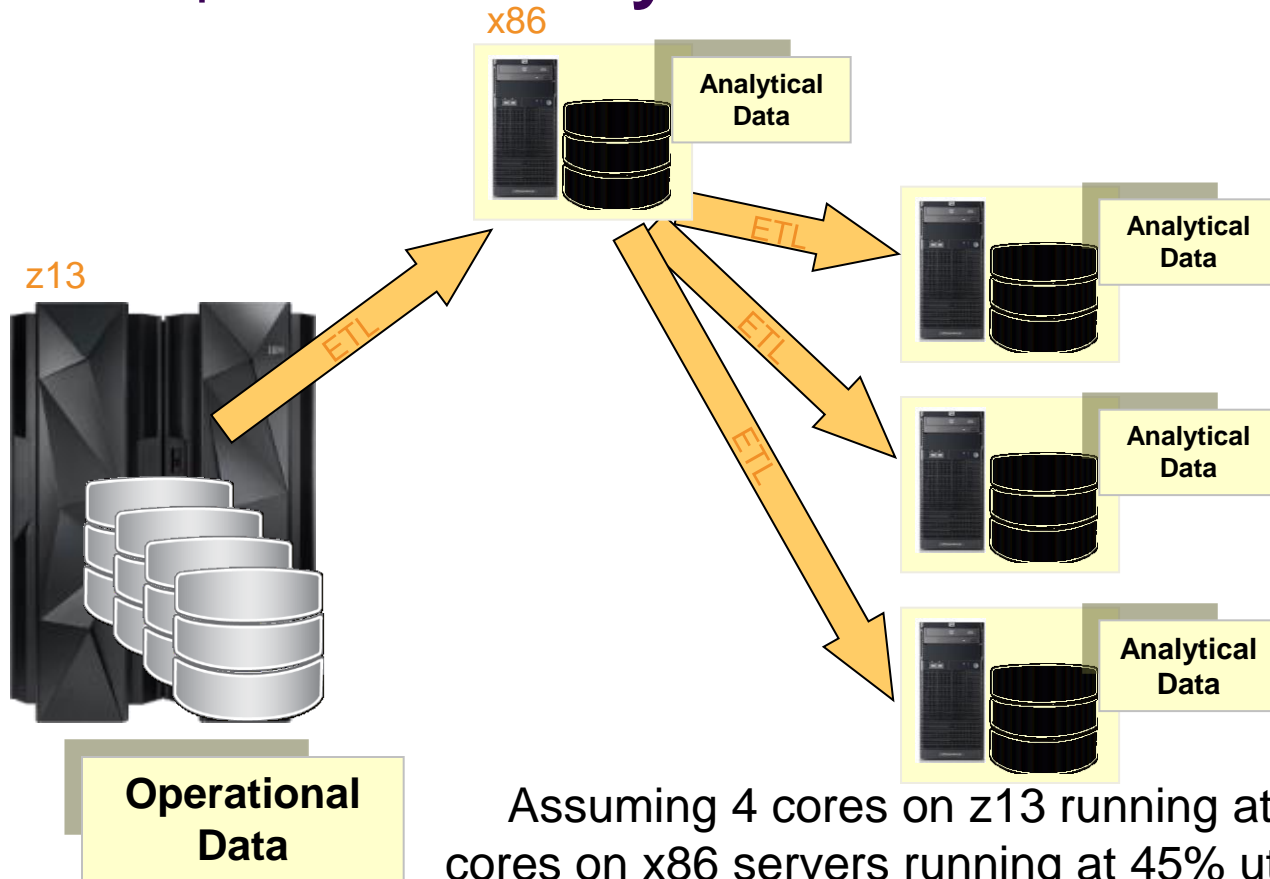
x86 and public cloud yield lower 3yr TCO

Case Study: 24 Workloads (24 VMs)



System configurations are based on equivalence ratios derived from IBM internal studies. Prices used are published US list prices as of 1/1/2015 for both IBM and competitors. Public cloud case includes costs of infrastructure (instances, data out, storage, free tier/reserved tier discounts), middleware and labor. z13 and x86 cases include costs of infrastructure (system, memory, storage, virtualization, OS, cloud mgmt), middleware, power, floor space and labor.

To move 1TB of data daily off z Systems can cost over \$10M over 4 years



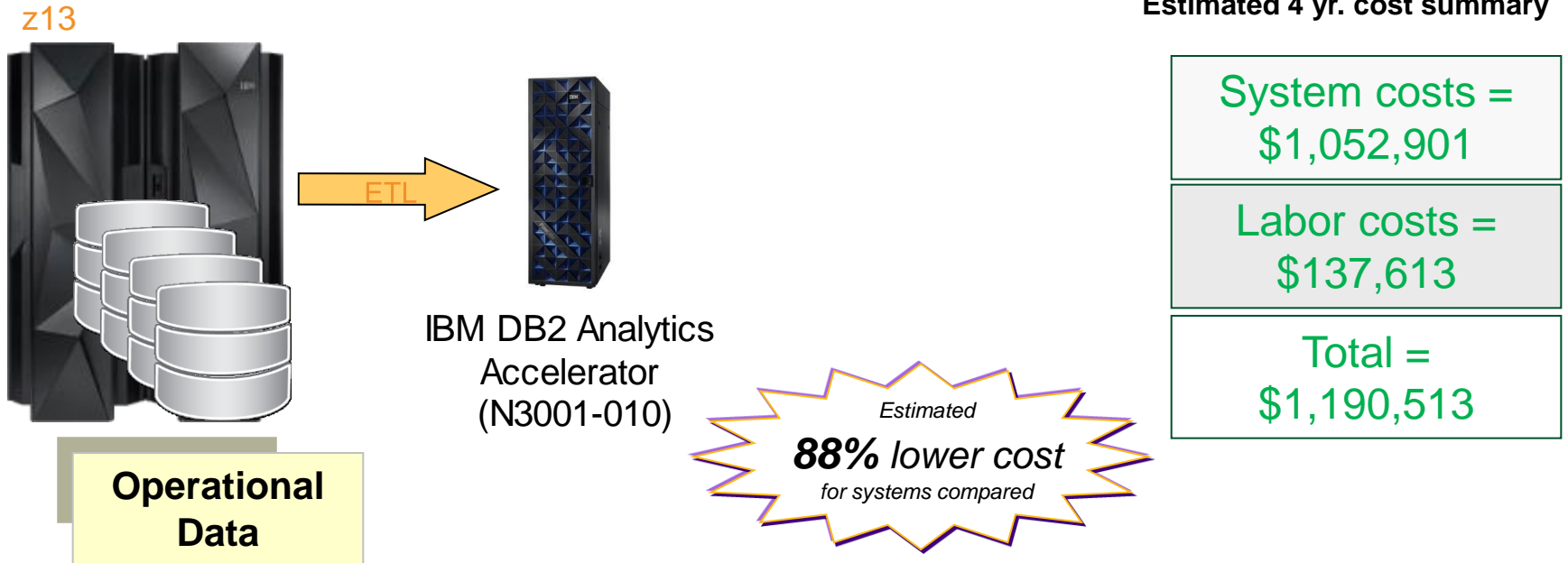
Estimated 4 yr. cost summary

System costs = \$9,864,412
Labor costs = \$393,927
Total = \$10,258,339

Assuming 4 cores on z13 running at 85% utilization and 12 cores on x86 servers running at 45% utilization, transfer will burn **519 MIPS** and use **10 x86 cores per day**

This is based on an IBM internal study designed to replicate a typical IBM customer workload usage in the marketplace. Test involved measuring in a controlled laboratory environment elapsed time for system and administrator to extract, send and receive 130GB file from z13 to an x86 server running with 12 x Xeon 2.4GHz E5-2440 processors. Prices, where applicable, are based on US prices as of 12/31/2014 for both IBM and competitor. Estimated amortized cost from 4 Year Total Cost of Acquisition (TCA) that includes all HW, SW (OS, DB and tools) and 4 years of service & support. For Labor costs, used annual burdened rate of \$159,600 for IT Administrator for z Systems and x86. Results may not be typical and will vary based on actual workload, configuration, applications, queries and other variables in a production environment. Users of this document should verify the applicable data for their specific environment.

Keeping the data on z13 and making a copy for DB2 Analytics Accelerator saves over 88%



Assuming 4 cores on z13 running at 85% utilization and 140 x86 cores on N3001-010 running at 45% utilization, transfer will burn **260 MIPS** and use **0.44 x86 core per day**

This is based on an IBM internal study designed to replicate a typical IBM customer workload usage in the marketplace. Test involved measuring in a controlled laboratory environment elapsed time for system and administrator to extract, send and receive 1,118GB file from z13 to DB2 Analytics Accelerator N3001-010 (Mako Full Rack). Prices, where applicable, are based on US prices as of 12/31/2014 for both IBM and competitor. Estimated amortized cost from 4 Year Total Cost of Acquisition (TCA) that includes all HW, SW (OS, DB and tools) and 4 years of service & support. For Labor costs, used annual burdened rate of \$159,600 for IT Administrator for z Systems and x86. Results may not be typical and will vary based on actual workload, configuration, applications, queries and other variables in a production environment. Users of this document should verify the applicable data for their specific environment.

z Systems Is Optimized For Operational Analytics

17x performance
13x price performance!
for systems compared

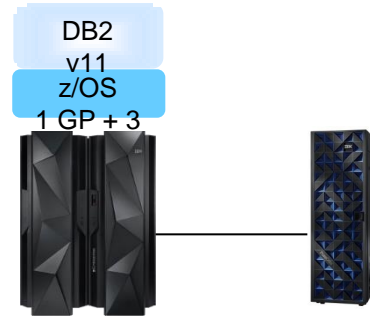
**Standalone
Pre-integrated
Competitor V4**



**Eigth
Unit**

Workload Time	1,810 mins
Reports per Hour (RpH)	5,343
Competitor Eigth Unit (HW+SW+Storage)	\$2,746,041

\$514
 Per Report per Hour
 (3yr TCA at no discount)



z13

**IBM DB2 Analytics
Accelerator
(N3001-010)**

Workload Time	105 mins
Reports per Hour (RpH)	92,095
z13 (1 GP + 3 zIIP, HW+SW+ Storage) + Accelerator V4.1 with PDA N3001-010 hardware	\$3,652,131

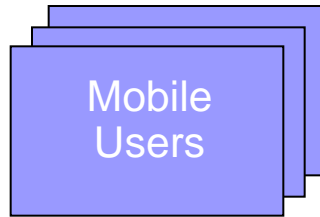
**IBM zEnterprise
Analytics System
9700**

\$40
 Per Report per Hour
 (3yr TCA at no discount)

Based on IBM sponsored and internal tests comparing IBM zEnterprise Analytics System 9700 with a comparably priced, comparably tuned competitor Eigth Unit configuration (version available as of 12/31/2014), executing a materially identical 10 TB BIDAY "Fixed Execution" workload in a controlled laboratory environment. Test conducted with BIDAY "Fixed Execution" workload measures elapsed time for executing 161,166 concurrent reports using 80 concurrent users. Intermediate and complex reports are automatically redirected to IBM DB2 Analytics Accelerator for z/OS (powered by N3001-010 hardware or Mako). Price comparison based on a 3YR Total Cost of Acquisition (TCA) using U.S. prices current as of December 31, 2014, including hardware, software, and maintenance. Compared prices exclude applicable taxes, and are subject to change without notice. Competitor configuration: Eigth Unit including competitor recommended software options and features. IBM configuration: z13 platform with 1CP and 3 zIIPs with 128GB memory and DB2 Analytics Accelerator Full Rack (N3001-10) with 7 S-blades (140 Intel E5-2680v2 2.8GHz cores and 128 GB RAM), 2 Hosts (1 active – 1 passive) with 20 Intel E5-4650v2 2.4GHz cores each and 12 disk enclosures, each with 24 600GB SAS drives. Results may not be typical and will vary based on actual workload, configuration, applications, queries and other variables in a production environment. Users of this document should verify the applicable data for their specific environment.

Oracle Coherence reduces TCA for read-only severe *sticky finger* with *think-time* user mobile workloads by 57% (forcing cache update)

Which platform provides the lowest TCA over 3 years?

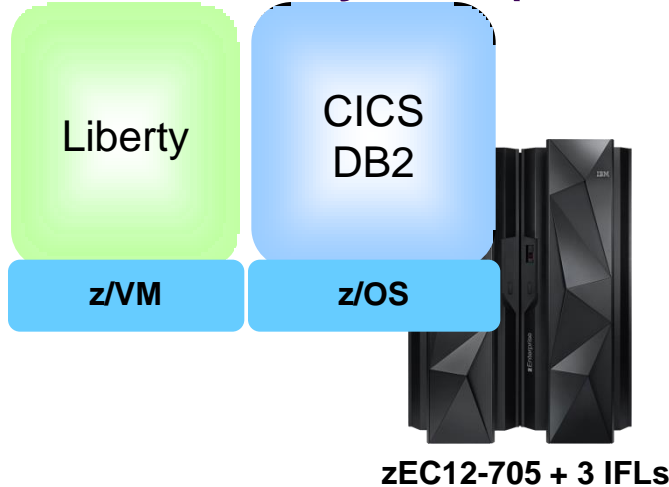


- 500 concurrent connections
- 20 reads/session with 100ms think time (forcing a cache refresh)
- 1 second cache invalidation (WXS scenario)

Mobile read-only workload driving minimum throughput of **5,200** transactions per second and response time of 5ms

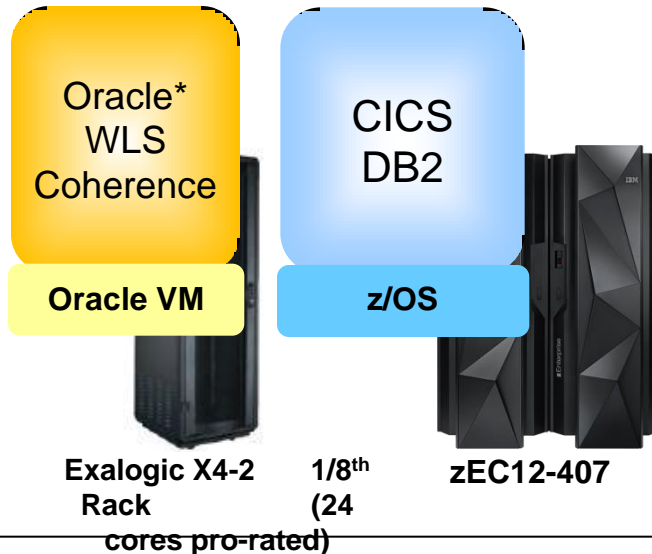
* Oracle Coherence performance projected from WXS Caching Test

WXS caching study for mobile workload - IBM Confidential



\$21.8M (3 yr. TCA)
Prod

\$28.5M (3 yr. TCA)
Prod+Dev/QA+DR



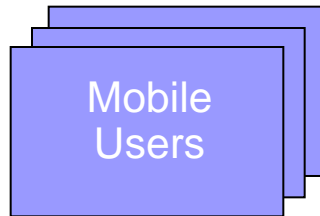
\$8.6M (3 yr. TCA)
Prod

\$12.3M (3 yr. TCA)
Prod+Dev/QA+DR

57%
lower cost!

Oracle Coherence increases TCA by 5% for read-only moderate sticky finger with think-time user mobile workloads (forcing cache update) – using Mobile Workload Pricing

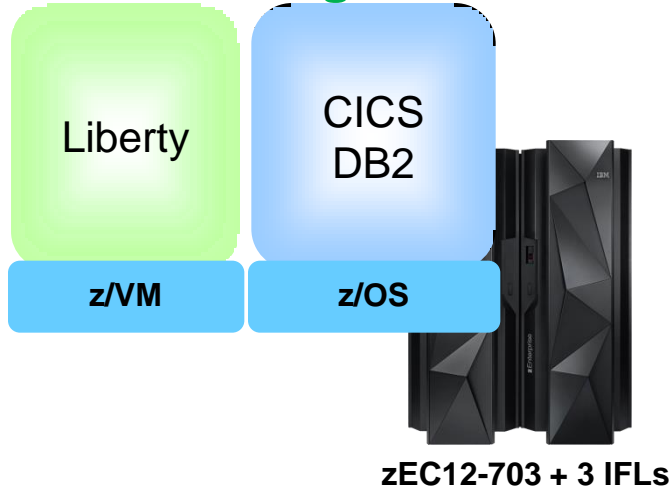
Which platform provides the lowest TCA over 3 years?



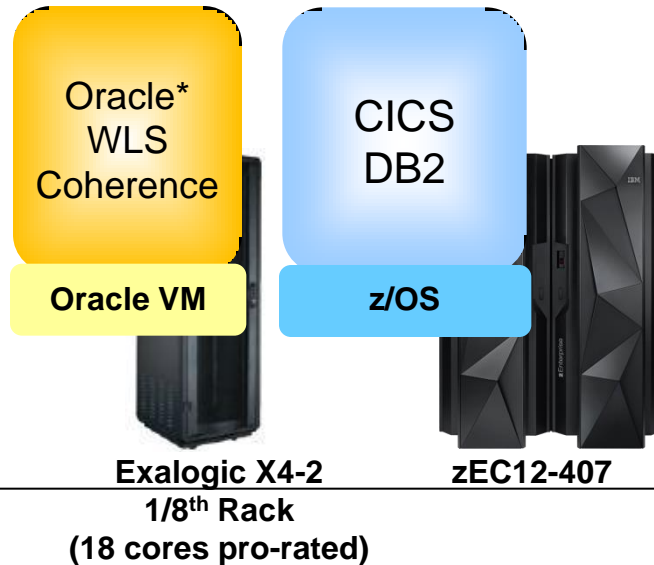
- 500 concurrent connections
- 10 reads/session with 200ms think time (forcing a cache refresh)
- 1 second cache invalidation (WXS scenario)

Mobile read-only workload driving minimum throughput of 3400 transactions per second and response time of 2ms

* Oracle Coherence performance projected from WXS Caching Test



\$8.5M (3 yr. TCA)
Prod
\$11.2M (3 yr. TCA)
Prod+Dev/QA+DR

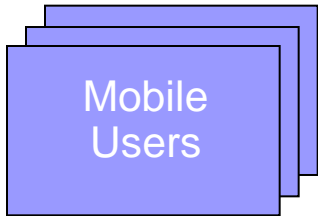


\$8.4M (3 yr. TCA)
Prod
\$11.8M (3 yr. TCA)
Prod+Dev/QA+DR

5%
higher cost!

Replicating z Systems Mobile Workloads increases TCA by 66% versus co-locating MobileFirst Platform and using Mobile Workload Pricing

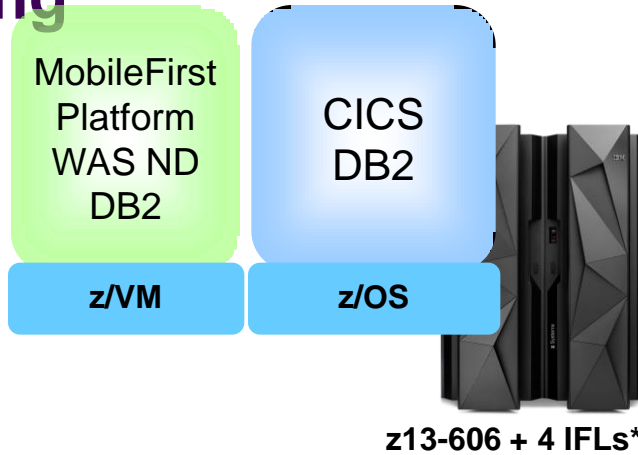
Which platform provides the lowest TCA over 3 years?



- 500 concurrent connections
- 70% do 1 read/session; 25% do 4 reads/session; 5% do 20 reads/session with 100ms think time
- 1 second cache invalidation

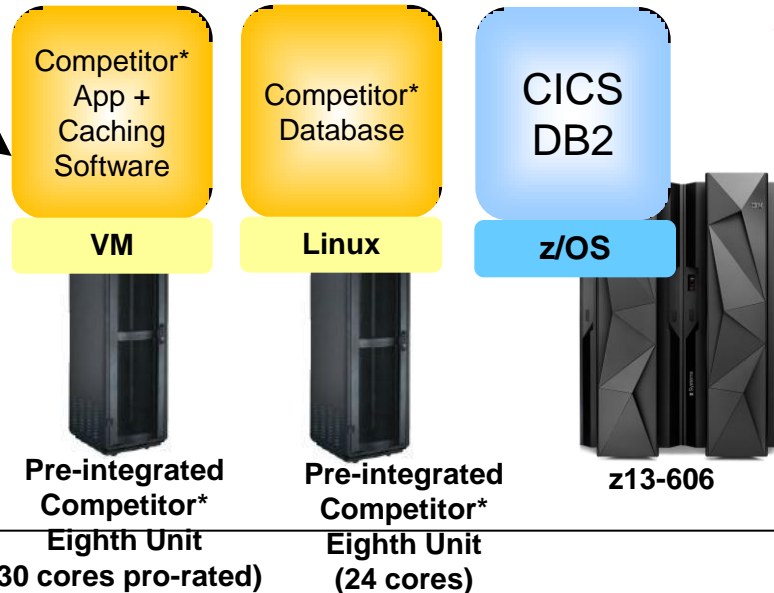
Mobile read-only workload driving minimum throughput of **6,300** transaction per second and response time of 12ms

* Competitor Caching and Database sizing estimated from WebSphere Extreme Scale Caching Test.
 ** Estimated performance, sizing and cost for z13 based on tests conducted on zEC12



\$11.2M (3 yr. TCA)
 Prod + Dev/QA + DR
 Mobile Workload Pricing

Estimated **40%** lower cost for systems compared



\$18.6M (3 yr. TCA) Prod + Dev/QA + DR


Estimated **66%** higher cost for systems compared

This is based on an IBM internal study designed to replicate a typical IBM customer workload usage in the marketplace. Test involved executing a materially identical mobile transaction processing workload in a controlled laboratory environment with comparable tuning and sizing. Prices, where applicable, are based on US prices as of 12/31/2014 for both z13 and competitor. Price comparison based on 3 Year Total Cost of Acquisition (TCA) includes all HW, SW and 3 years of service & support. Sizing shown is for Production to which 30% is added for System z for Dev/QA and CBU pricing for DR and 2x for Distributed.

MobileFirst Platform on Linux on z System* is expected to provide lower front-end cost and better scalability than x86

	MobileFirst Platform on Linux on z Systems*		MobileFirst Platform on x86	
# Concurrent Users	Front-end Cost per TPS	Response Time (ms)	Front-end Cost per TPS	Response time (ms)
10	\$2,634	42	\$2,074	50
30	\$1,091	43	\$1,066	54
50	\$812	44	\$964	62
100	\$525	48	\$770	68
200	\$456	70	\$636	95
400	\$439	131	\$693	205

At 50 concurrent users, z Systems provides better 3-year TCA



16% better
↓
37% better

Green = Better

* Estimated performance, sizing and cost for z13 based on tests conducted on zEC12

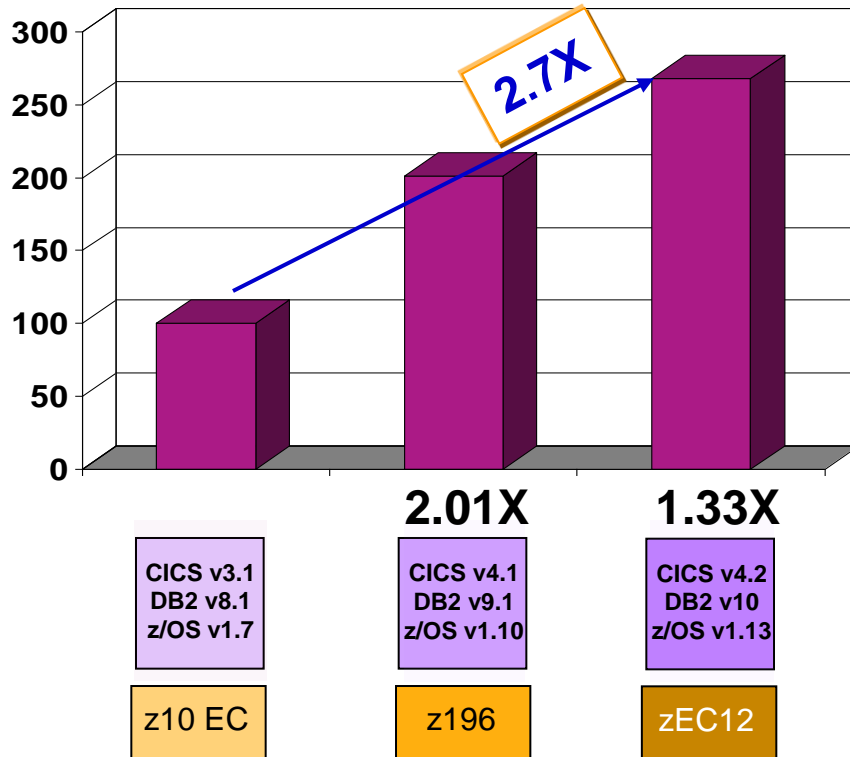
This is based on an IBM internal study designed to replicate a typical IBM customer workload usage in the marketplace. Test involved measuring throughput in transactions per second and response time for executing a materially identical mobile transaction processing workload in a controlled laboratory environment with comparable tuning and sizing. Prices, where applicable, are based on US prices as of 12/31/2014 for both IBM and competitor. Price comparison based on 3 Year Total Cost of Acquisition (TCA) includes all HW, SW and 3 years of service & support. Sizing shown is for Production to which 30% is added for System z for Dev/QA and CBU pricing for DR and 2x for Distributed.

Lessons Learned Can Be Grouped Into Three Broad Categories

- Always compare to an optimum z System environment
- Look for not-so-obvious distributed platform costs to avoid
- Consider additional platform differences that affect cost



Performance Improvements Can Lower MLC Costs And Free Up Hardware Capacity



IBM internal core banking workload (Friendly Bank). Results may vary.

Customer examples:

(1) Large MEA bank

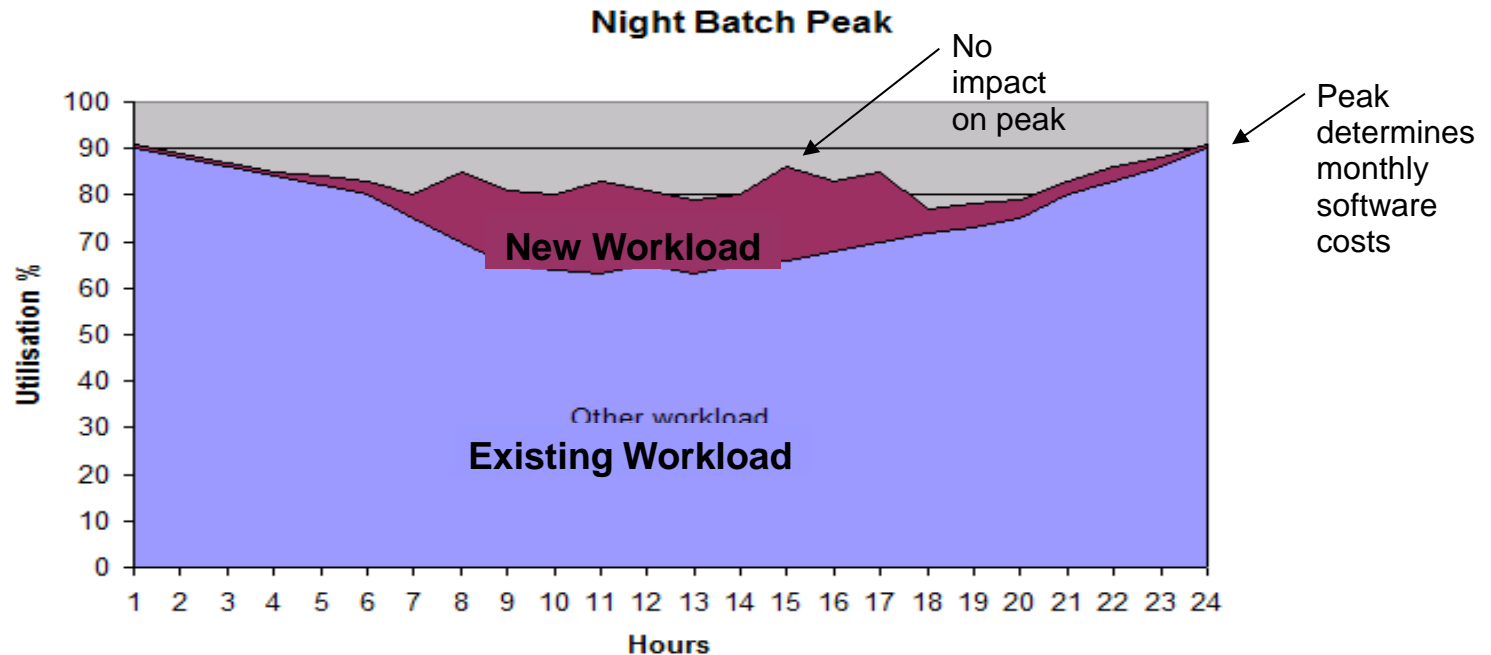
- Delayed upgrade from z/OS 1.6 because of cost concerns
- When finally did upgrade to z/OS 1.8
 - ▶ Reduced each LPAR's MIPS by 5%
 - ▶ Monthly software cost savings paid for the upgrade almost immediately

(2) Large European Auto company

- Upgraded to DB2 10
- Realized 38% pathlength reduction for their heavy insert workload
 - ▶ Other DB2 10 users saw 5-10% CPU reduction for traditional workloads

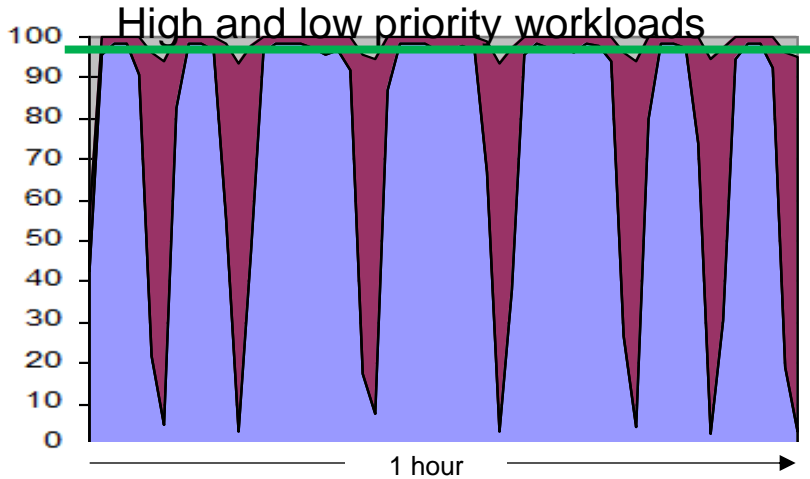
Additionally, save costs by moving to newer compilers and tuning

Sub-Capacity May Produce Free Workloads



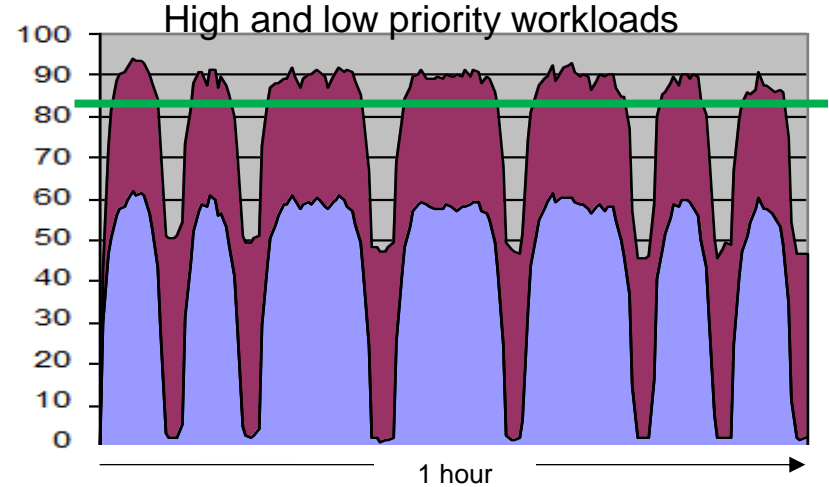
- Standard “overnight batch peak” profile – drives monthly software costs
- Hardware and software are free for new workloads using the same middleware (e.g. DB2, CICS, IMS, WAS, etc.)
- Ensure you exploit any free workload opportunities, and conversely, avoid offloading free applications!

z Systems has advanced workload management, guaranteeing service delivery and high utilization



z Systems –
Advanced workload management

High priority workloads (blue) run at very high utilization and do not degrade
 Low priority workloads (maroon) consume all but 2% of remaining resources (gray)



x86 hypervisor –
Imperfect workload management

High priority workloads (blue) run at *less* high utilization and *degrade* when low priority workloads (maroon) *added*
 Too much resource (gray) *remains unused* (22%)

Disaster Recovery On z System Costs Much Less Than On Distributed Servers

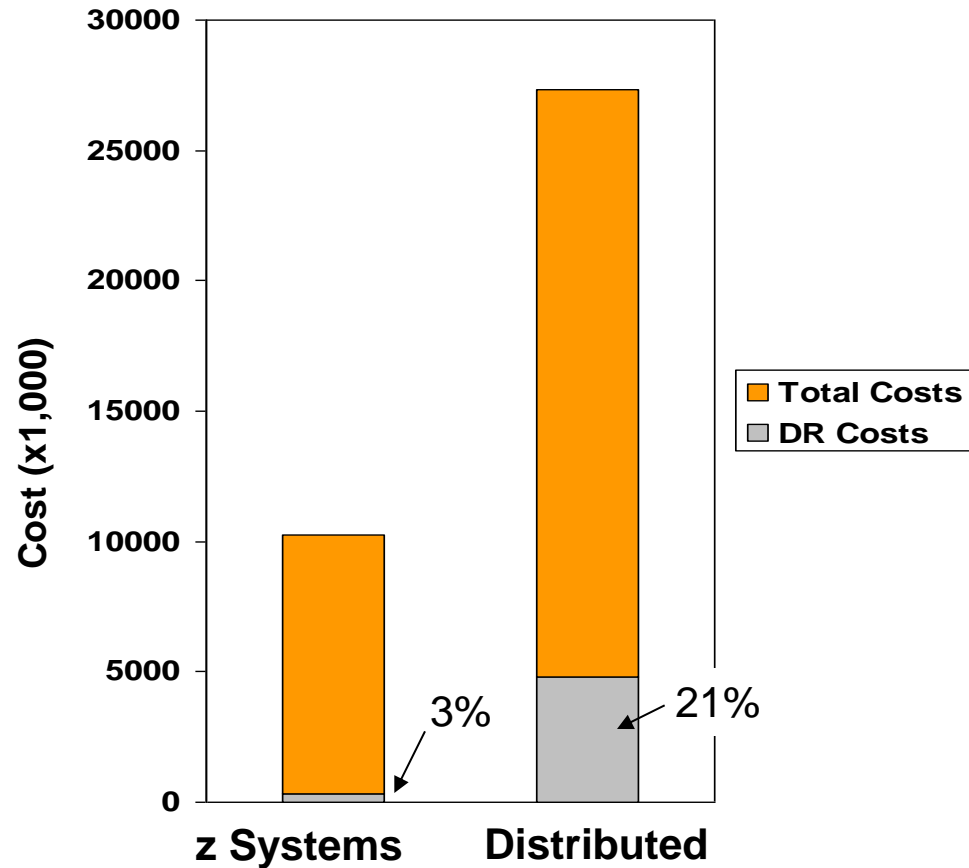
A large European insurance company with mixed distributed and z Systems environment at :

Disaster Recovery Cost as a percentage of Total Direct Costs:

z Systems— **3%**

Distributed – **21%**

Two mission-critical workloads on distributed servers had DR cost > 40% of total costs



Disaster Recovery Testing Is Typically More Expensive On Distributed Platforms Too

- A major US hotel chain
 - ~ 200 Distributed Servers (LinTel, Wintel, AIX, and HP-UX)

	<i>Person-hours</i>	<i>Elapsed days</i>	<i>Labor Cost</i>
<i>Infrastructure Test (7 times)</i>	1,144	7	\$89,539
<i>Full Test (4 times)</i>	2,880	13	\$225,416
Annual Total – Distributed	14,952*	73	\$1,170,281
Mainframe Estimate	2,051*	10	\$160,000

* Does not include DR planning and post-test debriefing

- Customer Recovery Time Objective (RTO) estimates:
 - Distributed ~ 48 hours to 60 hours
 - Mainframe ~ 2 hours
- Conclusion: Mainframe both simplifies and improves DR testing

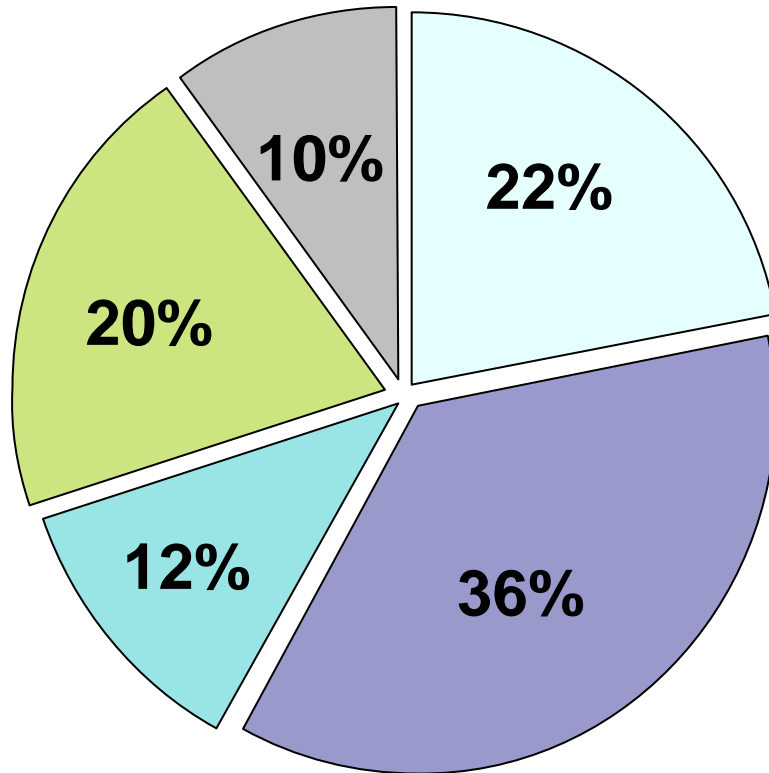
Accumulated Field Data For Labor Costs

- Average of quoted infrastructure labor costs
 - **30.7** servers per FTE (dedicated Intel servers)
 - **67.8** hours per year per server for hardware and software tasks
 - **52.5** Virtual Machines per FTE (virtualized Intel servers)
 - **39.6** hours per year per Virtual Machine for software tasks and amortized hardware tasks
 - Typical 8 Virtual Machines per physical server

- Best fit data indicates
 - Hardware tasks are **32** hours per physical server per year
 - Assume this applies to Intel or Power servers
 - Internal IBM studies estimate **320** hours per IFL for zLinux scenarios
 - Software tasks are **36** hours per software image per year
 - Assume this applies to all distributed and zLinux software images

Five Key IT Processes For Infrastructure Administration

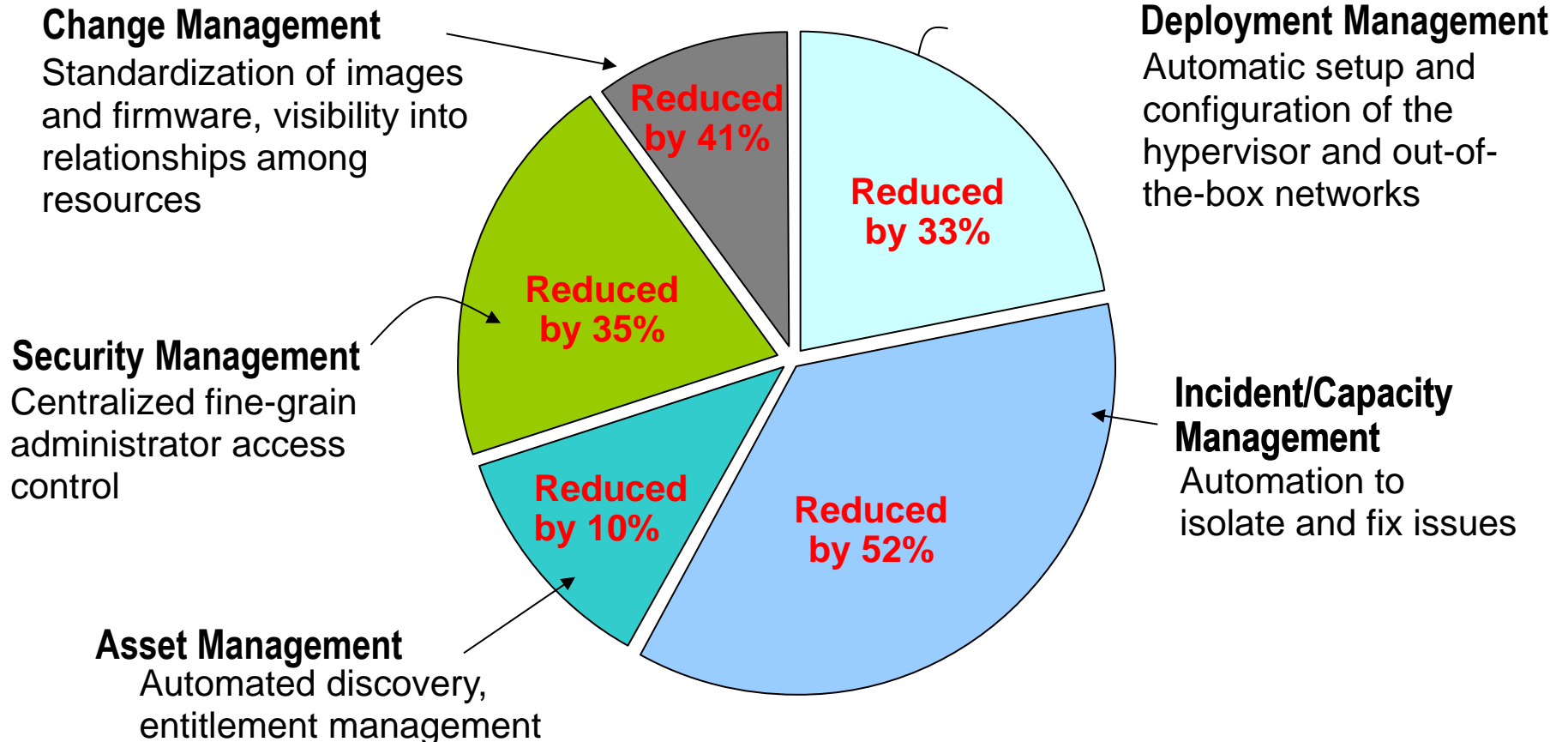
Time spent on each activity



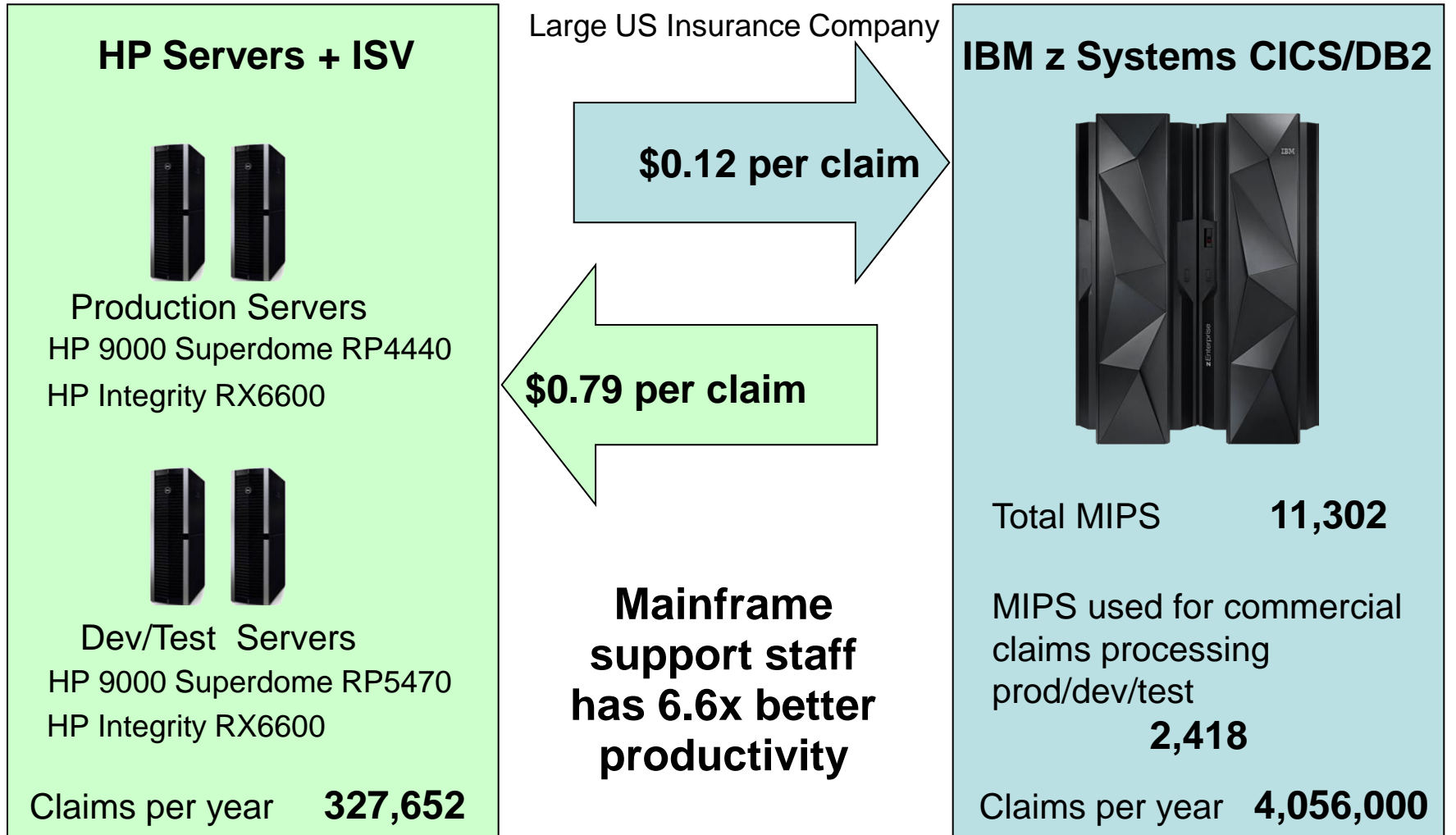
- Deployment Management**
– Hardware set-up and software deployment
- Incident/Capacity Management**
– Monitor and respond automatically
- Asset Management**
– Hardware and software asset tracking
- Security Management**
– Access control
- Change Management**
– Hardware and software changes

Z System Labor Cost Reduction Benefits Case Study

5032 total hours per year **reduced by 38%** to 3111 hours per year



Large Systems With Centralized Management Deliver Better Labor Productivity



TCO: Understand The Complete Picture



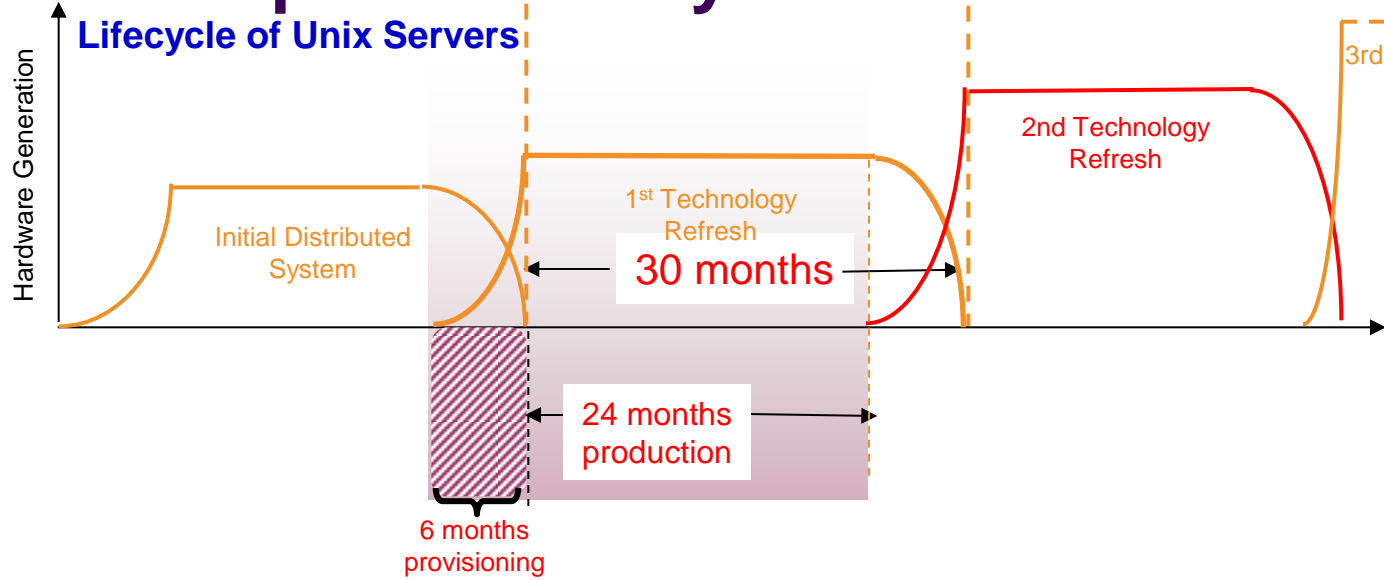
Thank you.

Cost Ratios in all TCO Studies

Average Cost Ratios (z vs Distributed)

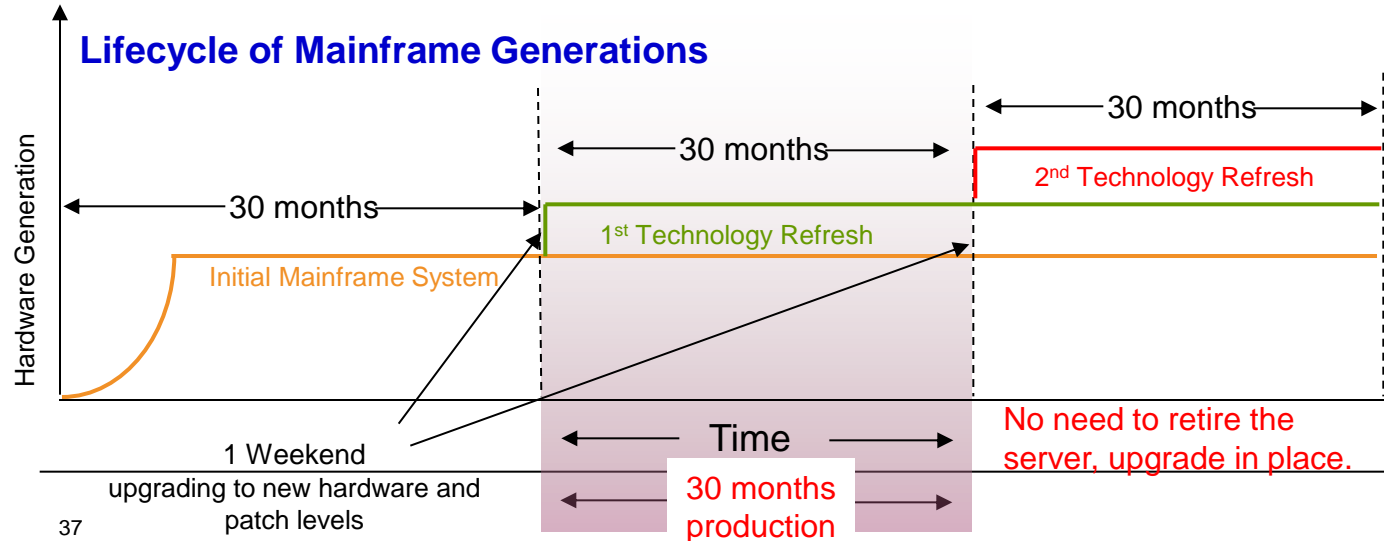
		z	Distributed	z vs distributed (%)
Offload	5-Year TCO	\$16,351,122	\$31,916,262	51.23%
	Annual Operating Cost	\$2,998,951	\$4,405,510	68.07%
	Software	\$10,932,610	\$16,694,413	65.49%
	Hardware	\$3,124,013	\$3,732,322	83.70%
	System Support Labor	\$3,257,810	\$4,429,166	73.55%
	Electricity	\$45,435	\$206,930	21.96%
	Space	\$59,199	\$154,065	38.42%
	Migration	\$438,082	\$10,690,382	4.10%
	DR	\$854,266	\$2,683,652	31.83%
	Average MIPS	3,954		
	Total MIPS	217,452		
Consolidation	5-Year TCO	\$5,896,809	\$10,371,020	56.86%
	Annual Operating Cost	\$716,184	\$1,646,252	43.50%
	Software	\$2,240,067	\$6,689,261	33.49%
	Hardware	\$2,150,371	\$1,052,925	204.23%
	System Support Labor	\$1,766,403	\$2,395,693	73.73%
	Electricity	\$129,249	\$365,793	35.33%
	Space	\$84,033	\$205,860	40.82%
	Migration	\$678,449	\$0	
	DR	\$354,735	\$411,408	86.22%
	Average MIPS	10,821		
	Total MIPS	292,165		

Distributed Servers Need To Be Replaced Every 3 To 5 Years



Refresh is normally even worse than just re-purchasing existing capacity as this real customer demonstrates:

Non-mainframe systems must co-exist for months at a time while being refreshed, requiring space, power, licenses etc. In this case only 24 months of productive work is realized for each 30 month lease period and the leases overlap up to 6 months



The mainframe by contrast is upgraded over a weekend and is fully productive at all times