Enterprise IT

Enterprise Grade requirements for IT Solutions deployment and related economic considerations

Lessons to Software Engineering students at University of Rome "Tor Vergata"

By Giuseppe Calavaro, Ph.D.
IT Economics Consulting
IBM

Disclaimer: The views and information herein are those of the presenter; they do not necessarily reflect the views of IBM.

Schedule

- Lesson: Enterprise IT and Non Functional Requirements
  - March 10th @ 11:30
- Lesson: Enterprise IT and Financial Implications
  - March 17th @ 11:30
- Presentation of a Case Study + Homework Assignment
  - March 24th @ 11:30
  - Ten days to deliver the homework by publishing on GitHub (Private repository with me as contributor) and sending the link to: Calavaro@uniroma2.it by April 4th at 23:59
- Homework discussion
  - April 7th @ 11:30
  - Students who will present their homework will have a bonus for the exam score
    - Presentation made by max 3 slides and 5 minutes talk to summarize what has been major outcome and message they bring as experience for future
- Available to discuss the project, homework and general Q&A
  - April 16th @ 16:00
- Exam (Esonero)
  - April 21st @ 11:30

Goals

- Audience Background
  - While **Software Engineering students** have developed **good skills on software application development**, and related technologies such as Agile, DevOps, DataBases, OS, Distributed Computing etc.
  - This experience is often **limited to experimental projects,** developed in very small teams during university's courses.
- The **goals** of these lessons is to introduce the students to:
  - What are typical **Enterprise Level** requirements for IT Environments
  - Provide some ideas of the **order of magnitude of NFRs targets**
  - Introduce some **technical elements** to address these requirements
  - Introduce basic economic considerations and concepts such as Total Cost of Acquisition (**TCA**) vs Total Cost of Ownership (**TCO**)
  - Provide a view on how Enterprises are approaching digital transformations themes such as the **Hybrid Cloud**

What is an Enterprise?

- **Enterprises are generally large corporations/organizations** that
  - Manage thousands to millions+ clients,
  - Have hundreds or even thousands of employees,
  - Requires to be available 24x7
  - Operate in strictly regulated industry
  - Provide services under stringent Service Level Agreements (SLA)
  - Manage demand peaks (that could be 1000x more than average)
  - **Support service request can be millions to billions per unit of time**

  *Examples are: a Bank (**Finance**), an Insurance company (**Insurance**), a Large supermarket chain (**Retails**), an Airline company (**Transportation**), your Electricity company (**Utilities**), but also to Government agency dedicated to collect taxes such as the IRS (Public Administrations aka **PA**), a Large Hospitals (**Health**), etc.*

- "**Enterprise-level solutions**" are generally marketed as something that's very knowledge-intensive and requires significant investments
  - These solutions are implemented on a **large-scale basis** and require the attention of specialized IT technicians who understand how to implement them properly.

Deploying applications requires considering many aspects

Develop → Build → Run → Manage

Apps	Where You usually focus: **Application Lifecycle Management**		
Middle Ware	DB, Application Server, Comm Middleware	Data Replication, Partitioning	System Management, Tuning, Security, Data Masking etc.
Platform	OS, Virtualization, Containers	Orchestration, Routing Load Balancing,	Networking



What are the similarities and differences among this objects?

Where does Z fit in Hybrid Cloud

In the beginning there were applications...

Interaction with
the outside world

Cloud Apps
(**System of Engagement**)

Integration

Mainframe Apps
(**System of Record**)

Enterprise Technology,
Systems, Data, Network,
Security, Middleware, etc.

Other People's "stuff"

Content, Services, Security

Z stands for zero downtime and the IBM name for its own
Enterprise Computing Platform that is the evolution of Mainframes.

APIs: The critical elements to access all these services

Application Programming Interface (API)

- It is a set of programming instructions and standards for accessing software applications
  - Software-to-software interface
  - Also allow users to update data, companies to leverage core business processes and business logic
- Critical element in the development of mobile apps, which rely on back-end systems for support
  - Mainframe APIs expose the data that is requested from other APIs that are in the cloud or in other mobile applications
  - REST APIs provide great flexibility, allowing developers to build an API that will fulfil different needs

Purchase a Flight ticket Use Case

Get an airline reservation

Open the site or app of your favorite travel agent

Pick an airport, enter your travel dates and number of passengers

The app will display a list of possible flights from different airlines

Airlines public APIs provide access to their reservations systems

What are the enterprise concerns IT Engineers should know?

Non-Functional Requirements (NFRs)

- Reliability & Availability
- Serviceability
- Security
- Performances
- Scalability

Environments required

1. Development
2. Test
3. Pre-Production aka QA*
4. Production
5. High Availability
6. Disaster Recovery

** aka (Also Known As) QA (Quality Assurance)*

Reliability and Availability

- A 99% availability means 3 days and 15 hours of downtime per year
- A 99.999% availability means 5 mins of downtime per year.
- If there is a dependency, then availability is a product of the availability of the system and dependent system.
  - For example, if System A has 99.9% and System B has 99.9% and A depends on B, the availability of aggregate system is 99.9 * 99.9 = 99.8%
  - Very important concept when you have a system made of high number of components
- **Reliability**: component redundancy to allow continue deliver services in presence of faults.
- If there's redundancy of a system then the uptime is calculated as 100 minus product of redundant components failure rate.
  - For example, if the availability is 99.9, then failure rate is 0.1% and resulting availability is 100 minus (0.1 * 0.1) = 99.99% which shows an increased availability due to redundancy.
- **Six Sigma corresponds to 99.99966%** and works out to 1 minute 47 seconds of downtime per year

Availability %	Downtime per year[note 1]	Downtime per month	Downtime per week	Downtime per day
90% ("one nine")	**36.53 days**	**73.05 hours**	**16.80 hours**	**2.40 hours**
95% ("one and a half nines")	18.26 days	36.53 hours	8.40 hours	1.20 hours
97%	10.96 days	21.92 hours	5.04 hours	43.20 minutes
98%	7.31 days	14.61 hours	3.36 hours	28.80 minutes
99% ("two nines")	**3.65 days**	**7.31 hours**	**1.68 hours**	**14.40 minutes**
99.5% ("two and a half nines")	1.83 days	3.65 hours	50.40 minutes	7.20 minutes
99.8%	17.53 hours	87.66 minutes	20.16 minutes	2.88 minutes
99.9% ("three nines")	**8.77 hours**	**43.83 minutes**	**10.08 minutes**	**1.44 minutes**
99.95% ("three and a half nines")	4.38 hours	21.92 minutes	5.04 minutes	43.20 seconds
99.99% ("four nines")	**52.60 minutes**	**4.38 minutes**	**1.01 minutes**	**8.64 seconds**
99.995% ("four and a half nines")	26.30 minutes	2.19 minutes	30.24 seconds	4.32 seconds
99.999% ("five nines")	**5.26 minutes**	**26.30 seconds**	**6.05 seconds**	**864.00 milliseconds**
99.9999% ("six nines")	**31.56 seconds**	**2.63 seconds**	**604.80 milliseconds**	**86.40 milliseconds**
99.99999% ("seven nines")	**3.16 seconds**	**262.98 milliseconds**	**60.48 milliseconds**	**8.64 milliseconds**
99.999999% ("eight nines")	**315.58 milliseconds**	**26.30 milliseconds**	**6.05 milliseconds**	**864.00 microseconds**
99.9999999% ("nine nines")	**31.56 milliseconds**	**2.63 milliseconds**	**604.80 microseconds**	**86.40 microseconds**

Sources:
https://en.wikipedia.org/wiki/High_availability
https://searchunifiedcommunications.techtarget.com/tip/The-truth-about-five-nines-availability-in-unified-communications-networks

Reliability and Availability

- 24x7 with 99.999% availability means always ON.
  - It is not only a matter of **brand protection**
  - It is because **dictated by compliance rules** and regulations per industry
- Any Industry today has many regulations under which the Enterprises, belonging to such industry, must comply
  - As an example, banks are required to have **High Availability (HA)** and **Disaster Recovery (DR)** systems located in places hundred kilometer apart, to guarantee operations in case on natural disasters
- The **increase of number of systems leads to increase the number of failure events** that need to be handled
  - Achieving highest levels of availability means higher cost and complexity of systems in deployments/capacity addition/rollback etc.
- **Enterprise grade systems such as Z systems,** and its associated software, have evolved to the point that customers often experience months or even **years** of system availability between system downtimes.
  - Moreover, when the system is unavailable because of an unplanned failure or a scheduled upgrade, this period is typically very short

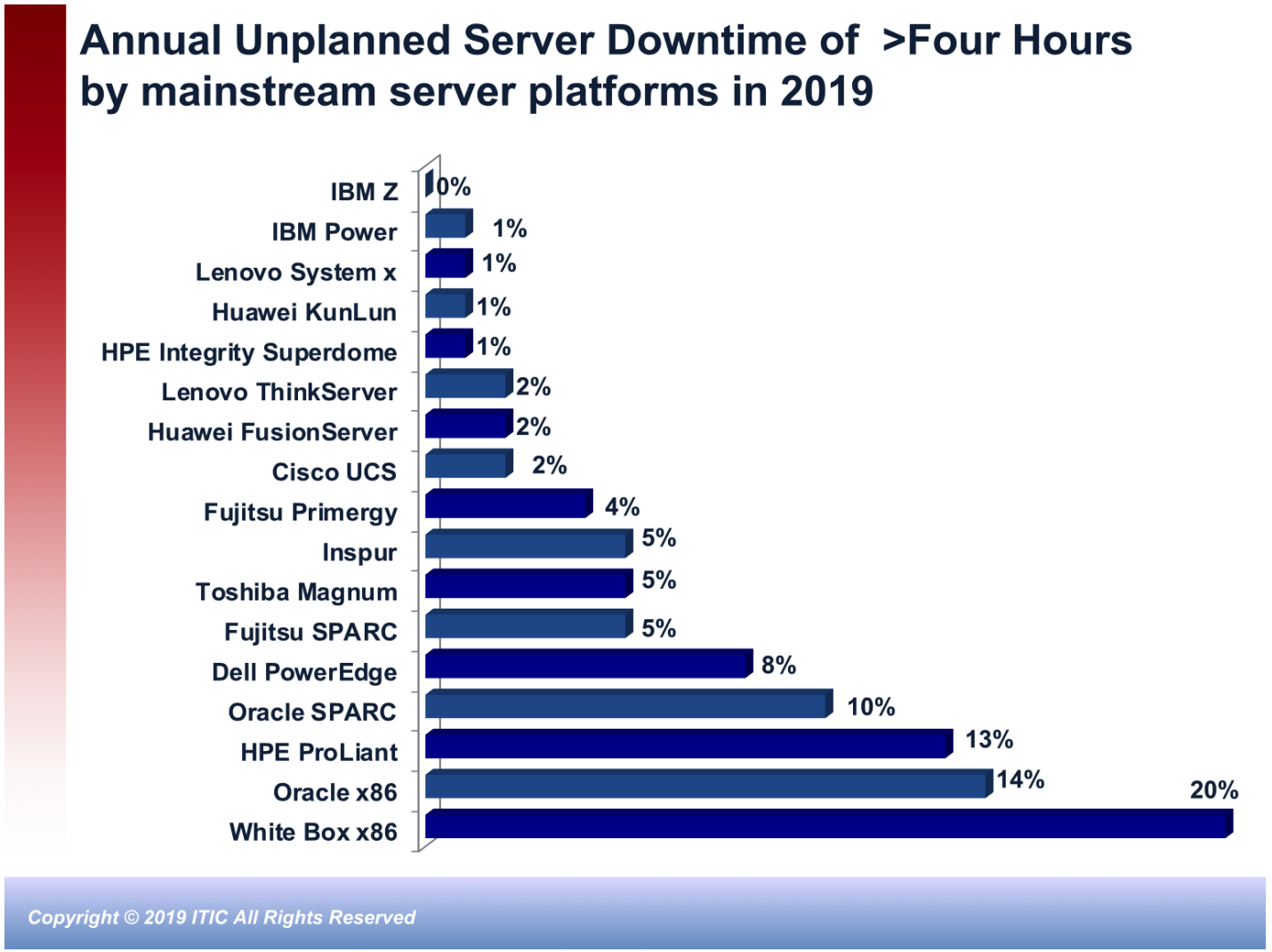
Server Reliability Rankings from Independent ITIC Survey

Also the September 2019 ITIC Server Reliability Survey (Mid-Year Update) confirmed the findings of the last years

ITIC's 2019 Reliability survey focus on:

- Server and OS reliability/uptime
- External factors impacting server reliability (e.g. security breaches; user carelessness)
- Top causes of unplanned downtime for servers, OS, virtualization & cloud
- Customer satisfaction with vendor service and support
- New questions on: Server and OS Availability and Downtime as a result of Server hardware component failures & timing of failures

Source: www.itic-corp.com



Copyright © 2019 ITIC All Rights Reserved

ITIC Survey results

- ITIC conducted separate surveys detailing the latest Annual Cost of Hourly Downtime; Downtime Costs by Vertical Market segments and Minimum Uptime and Reliability requirements and the impact of Security breaches
- The survey was independent: (No Vendor Sponsorship)
- Approximately 67% of respondents hailed from North America; 33% were international customers
- All market sectors represented: SMBs = 32%; SMEs = 26% and Enterprises = 43%
- Survey respondents hailed from 22 vertical markets
- ITIC used security & authentication to prevent tampering

Source: www.itic-corp.com

Survey Highlights: Reliability Trends

- ***Overall,*** the **inherent reliability** of the majority of server hardware, server operating systems and the underlying processor technology continues to improve. Human Error, Increasing complexity and Security issues undermine reliability particularly with respect to mainstream, "work horse" commodity servers.

- **Vendor Performance:**
  - **IBM Z mainframe is in a class of its own: best of breed reliability**
  - **IBM, Lenovo servers continued to deliver highest reliability over the last decade.**
  - **IBM, Lenovo, HPE Integrity, Huawei KunLun** record best Availability
  - **Cisco UCS reliability** notches dramatic reliability improvement in 2019 Mid-Year Update Poll as companies bolster, expand network edge resources and fortify security
  - **IBM and Lenovo server** reliability are up to **23x** more reliable than worst rivals
  - **HPE's** Integrity Superdome, **Stratus** ftServer and **Fujitsu** Primergy also scored high
  - **Lenovo, IBM, HPE and Huawei** rated highest in customer satisfaction

- **Reliability Trends:**
  - **Majority of corporations - 86% Require "Four Nines" of Uptime -** 99.99% for mission critical hardware, operating systems & main line of business (LOB) applications. This in an increase of four (4) percentage points from ITIC's 2017 – 2018 Reliability poll.
  - **Patch Time Increases:** 60% of firms now spend from two-to-four hours applying patches
  - **Increase in Server Workloads** causes reliability declines in 64% of servers >4 years old that haven't been retrofitted or upgraded to accommodate increased workloads.
  - **Cost of Hourly Downtime Increases: 98%** of firms say hourly downtime costs exceed $150K; **35%** of respondents estimate hourly downtime costs their companies up to $400K.

- **Top Issues Negatively impacting network reliability are:**
  - End User Carelessness – **74%;** Human Error (e.g., misconfiguration, right-sizing server workloads etc.) – **59%;** Security - **51%.**
  - Going forward Security issues will a persistent threat that can potentially undermine reliability

Serviceability (also known as supportability)

- **Serviceability** refers to the ability of technical support personnel to install, configure, and monitor computer products, identify exceptions or faults, debug or isolate faults to root cause analysis, and provide hardware or software maintenance in pursuit of solving a problem and restoring the product into service.
- The goal of serviceability is to **reduces operational costs and maintains business continuity**.

Examples of features that facilitate serviceability include:

- Help desk notification of exceptional events
- Network monitoring
- Documentation
- Event logging / Tracing (software)
- Logging of program state
- Software upgrade
- Graceful degradation
  - Where the product is designed to allow recovery from exceptional events without intervention by technical support staff
- Hardware replacement or software upgrade planning (**Concurrent Maintenance for both**)
  - Where the product is designed to allow efficient hardware upgrades with minimal computer system downtime (e.g., **hotswap components**.)

Security

Enterprise clients require extensive data protection ***due to compliance regulations and data breaches***

"It's no longer a matter of if, but when …"

28%

Likelihood of an organization having a data breach in the next 24 months [1]

Of the **14.7 Billion** records breached since 2013

only **4%** were encrypted [3]

$3.86M

Average cost of a data breach in 2018 [2]

72 records are stolen every second [3]

Compliance regulation have strict requirements

European Union GDPR

PCI-DSS

HIPAA

Encryption is one of the key elements for a Security Solution

Comprehensive data protection requires a huge investment.

It can be time consuming to deploy point solutions and/or enable encryption directly in every application used.

Organizations struggle with questions like these:

- *What data should be encrypted?*
- *Where should data encryption occur?*
- *Where should encrypted data reside?*
- *Who is responsible for encryption?*

Selective encryption challenges

Initial encryption approach:

"at-rest" protection for application and database data

- Identifying and classifying sensitive data can be barrier to deploying encryption
- Application-level encryption beyond small number of fields can be cost prohibitive
- Scope of data in flux due to new and changing regulatory mandates
- Processor consumption of application encryption at scale can be prohibitive
- Migration to encrypted data is disruptive
- Organizations are worried about privileged identities
- Full disk encryption alone may not meet regulatory requirements

Pervasive encryption … *A paradigm shift in data protection*

Protecting only enough data to achieve compliance should be the bare minimum, not a best practice

- Focus on eliminating barriers:
  - Decouple encryption from classification
  - Avoid extensive application changes
  - Protect databases *and* key files
  - Reduce the high cost associated with processor overhead

IBM Z Pervasive Encryption

Protect *all* application and database data according to enterprise security policies *without* interrupting business operations

IBM LinuxONE™

NO data classification

Bulk encryption enabled in the OS

- Simple implementation
- Transparent exploitation
- Optimized performance

NO processing impact to workloads

- Hardware-accelerated encryption on every core that is performed by dedicated unit that allow encryption with no impact on workload

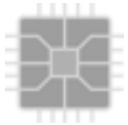
Secure Service Containers

These deliver

- Tamper-resistant installation and runtime
- Restricted administrator access
- Encryption of data and code

Pervasive Encryption with IBM Z

Enabled through tight platform integration

Integrated Crypto Hardware		Hardware accelerated encryption on every core, CPACF performance improvements of 7x Crypto Express6S – PCIe Hardware Security Module (HSM) & Cryptographic Coprocessor
Data at Rest		Broadly protect Linux file systems and z/OS data sets using policy controlled encryption that is transparent to applications and databases
Clustering		Protect z/OS Coupling Facility data end-to-end, using encryption that's transparent to applications
Network		Protect network traffic using standards based encryption from end to end, including encryption readiness technology to ensure that z/OS systems meet approved encryption criteria
Secure Service Container		Secure deployment of software appliances including tamper protection during installation and runtime, restricted administrator access, and encryption of data and code in-flight and at-rest
Key Management		The IBM Enterprise Key Management Foundation (EKMF) provides real-time, centralized secure management of keys and certificates with a variety of cryptographic devices and key stores

Performances

Computer performance is the amount of useful work accomplished by a computer system.

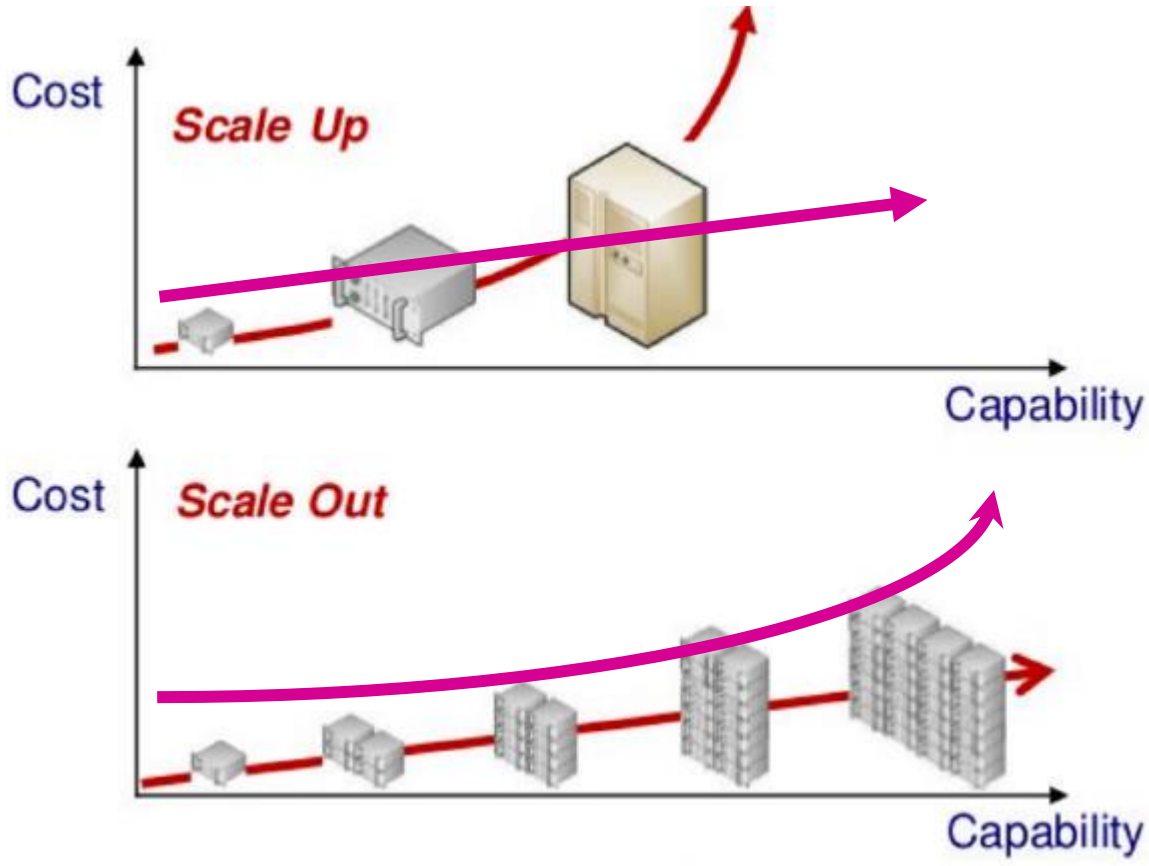
- Response time
- Processing Speed
- Channel Capacity
- Latency
- Bandwith
- Throughput
- Efficiency
- etc

What are the performance measures **for Systems of Records**?

- **Transactions per Seconds (TPS)**
- DB centric computation are at the heart of these systems
- Enterprises that use System of Records have stringent TPS requirements depending on their business size.
  - They must keep into account peaks and be able to deliver during peak season
  - Consider for example a retail business that knows what is the average number of transactions per second but knows as well that on black Friday his business must be able to work probably at x10 the average volume
  - Or consider a Credit Card system that must process the transaction while the payer is in the shop attempting to pay for his purchase at the register
- Examples:
  - Credit Cards purchase approval: up to millions of TPS
  - Tickets purchase: Hundred Thousands of TPS
  - Telephone calls logged: up to Millions of TPS

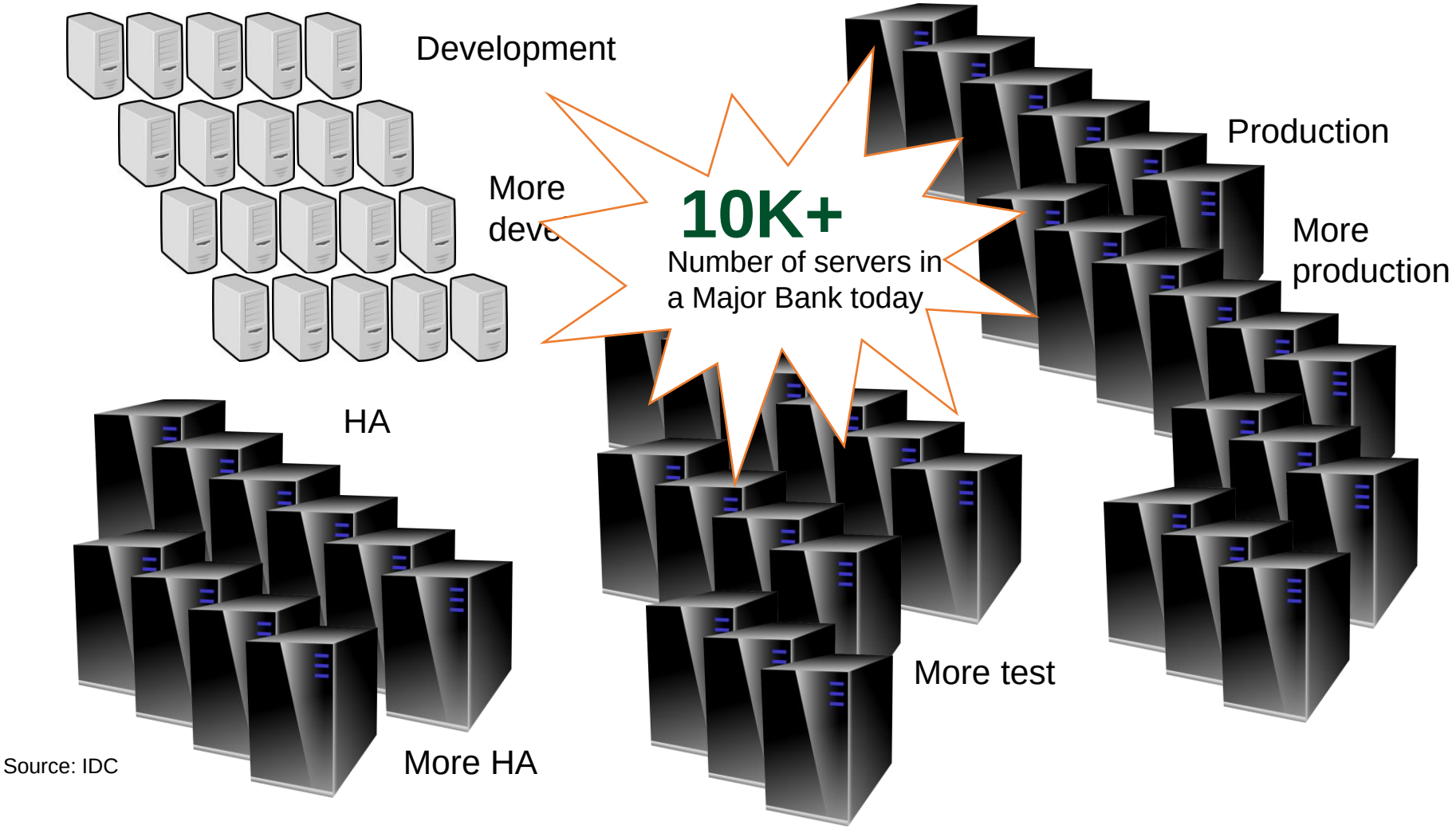
Scalability

- There are two major approaches to Scalability
  - **Scale Up** aka **Scale Vertically**
  - **Scale Out** aka **Scale Horizontally**
- It is common to find pictures, such as this one, that lead to think that Scaling out is cheaper than scaling up
  - Yet, it is Critical to consider all the costs for the Enterprise IT, when making such **TCO analysis**
  - This include also the costs of **Software**, **Networking**, **Space**, and **People**
    - For example, Databases are licensed by number of cores. If you increase x10 the number of cores, you will increase x10 the cost of software
- Provision for peak utilization lead to unused resources unless you can automatically reallocate these after peak

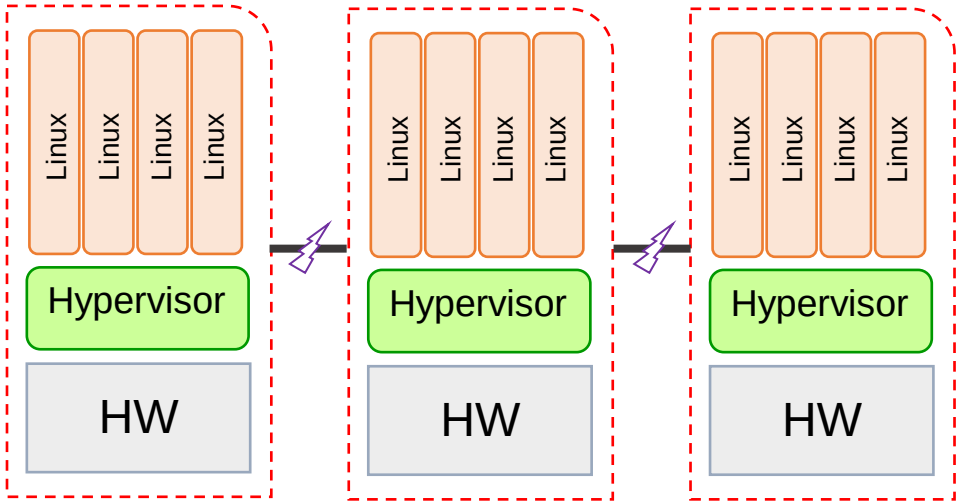


Scale-Out typical scenario

Continuing adding servers to data centers…



Most IT data centers are built around a scale-out model…



Blade mount

Rack mount

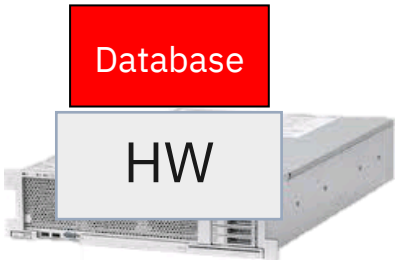
2017 to 2018

Rack Optimized and Blade servers growing @ 6%

Large systems growth @ 3%

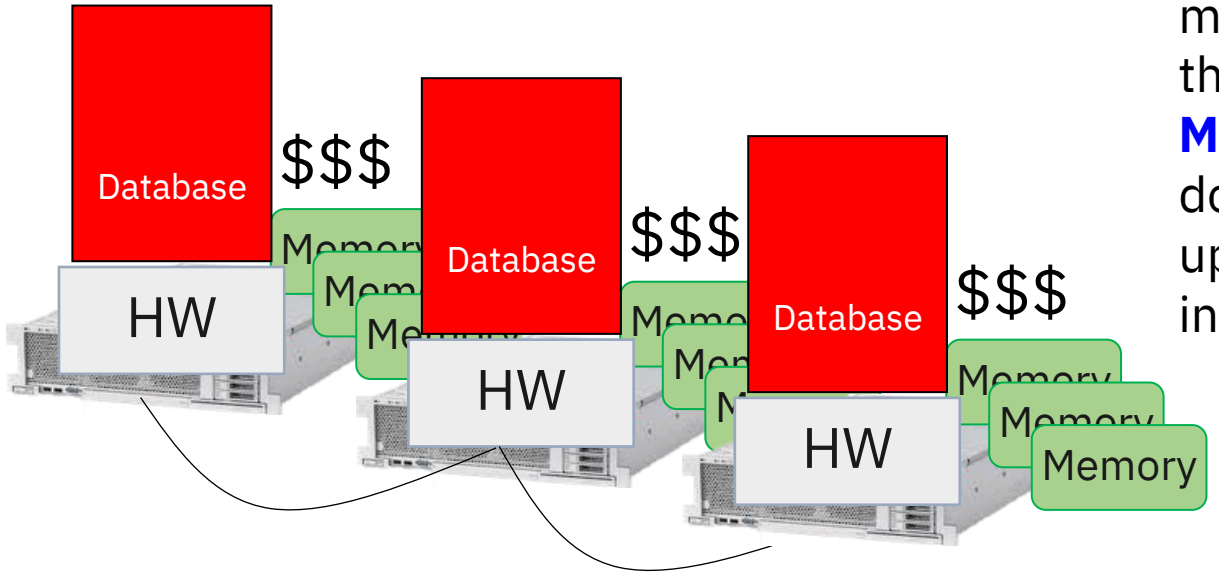
Source: IDC https://www.idc.com/getdoc.jsp?containerId=US42981218

Next to HW costs you must consider SW, Networks, etc.



Software costs are typically per Core or per server or per socket

Databases proliferate to support development, test, QA…



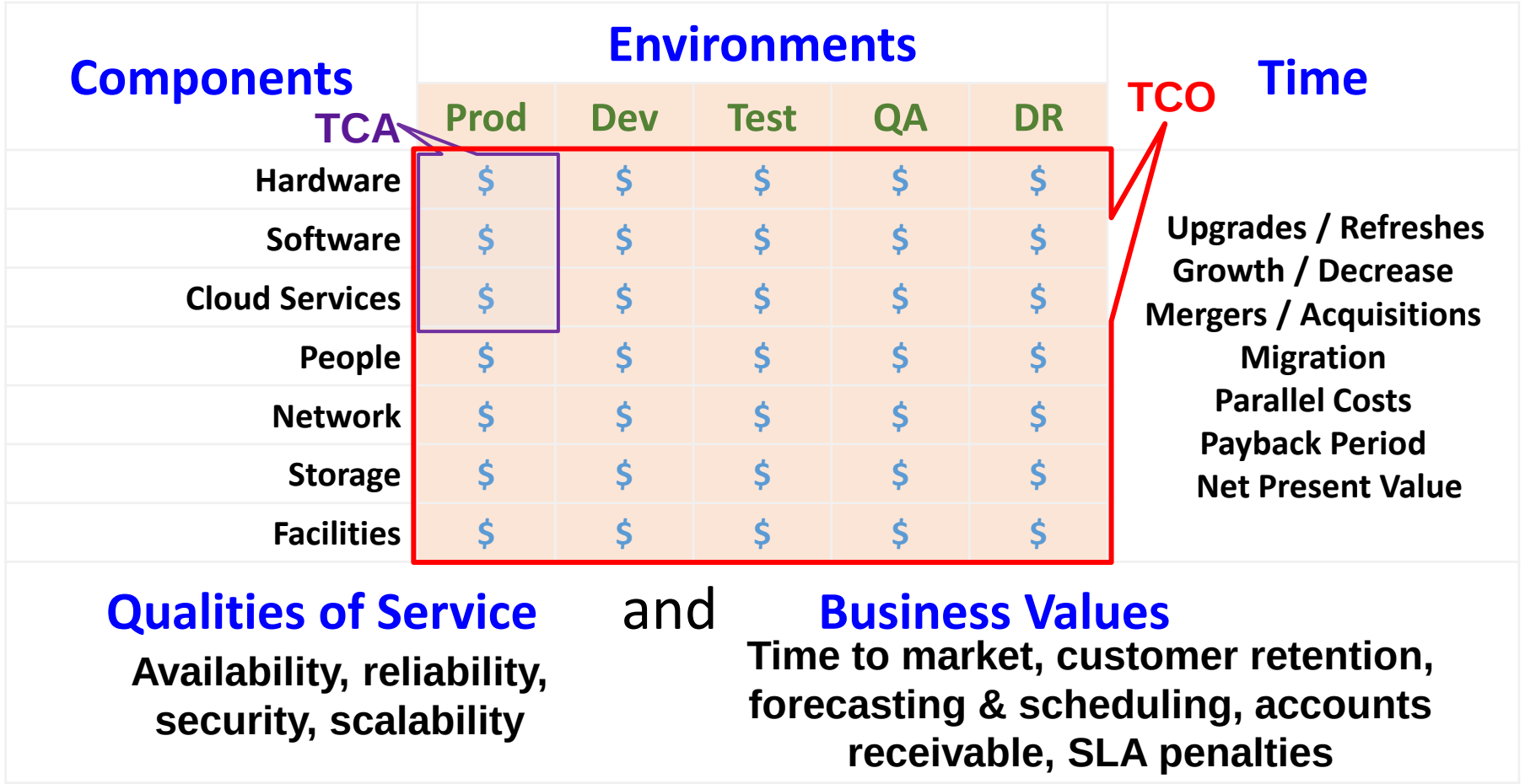
Most DB are licensed based on **total number of cores**

What is the cost of a DB per core for production Servers?

Also, you will need many more IT staffing people that will increase **Maintenance costs**: How do you manage the OS upgrades or patch installation?

TCO and Business Value – Get the complete picture

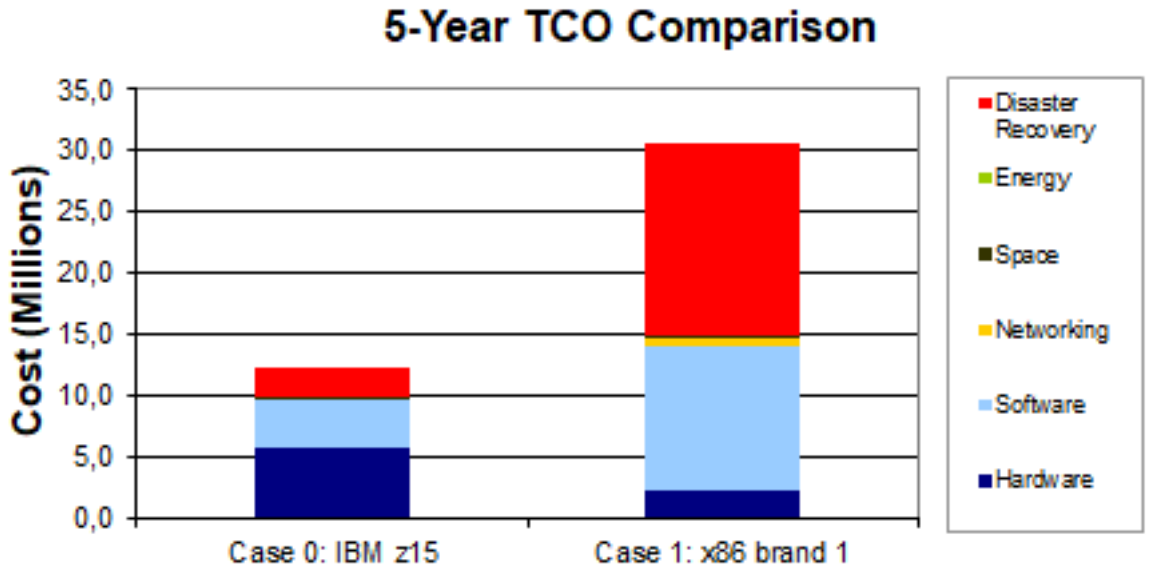
An **IT Economics study** quantifies business values, technical requirements and costs in a TCO



Total Cost of Ownership is much more than **Total Cost of Acquisition**!

IT Economics – Linux TCO Comparison on 5 years

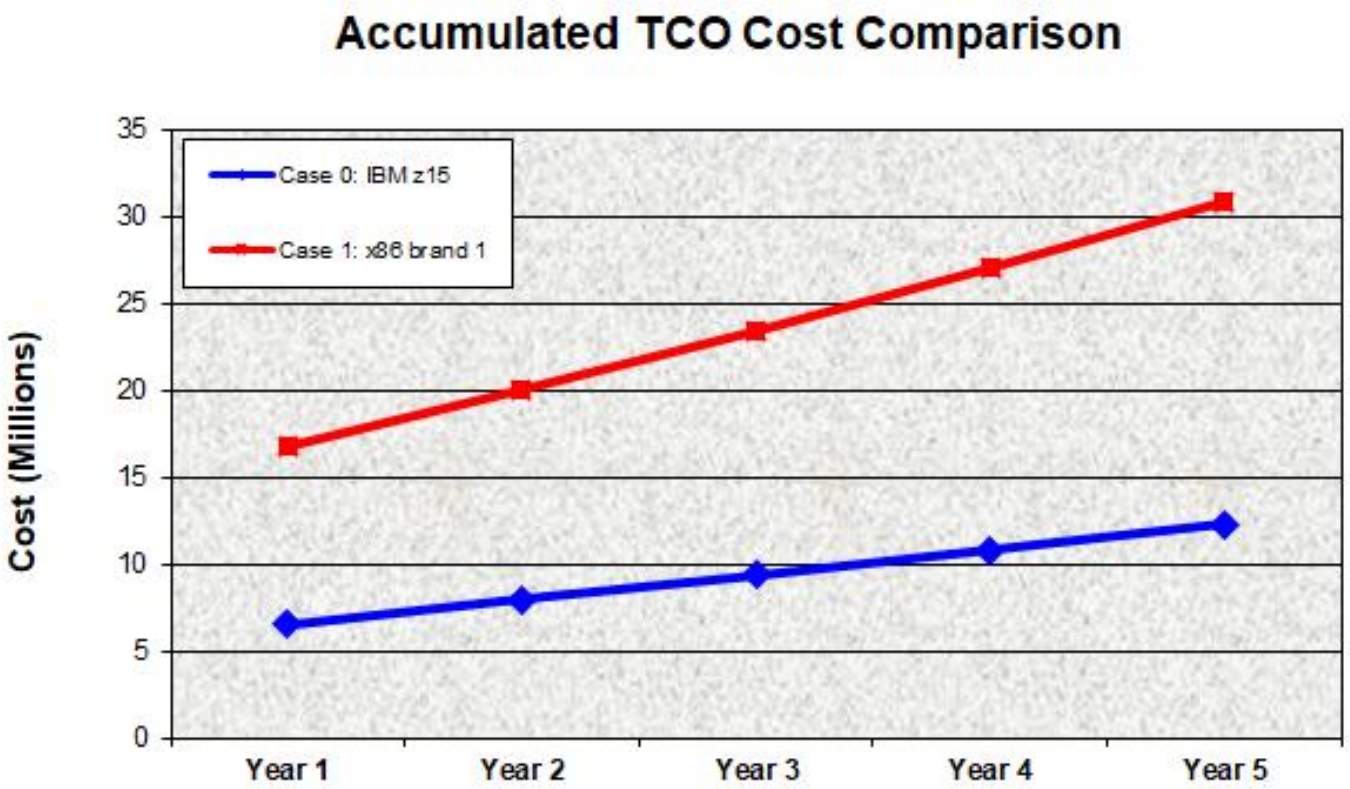
	Hardware	Software	Networking	Space	Energy	Disaster Recovery	TOTALS		Total # of Cores (Year 1)	# of servers
Case 0: IBM z15	5.733.646	3.887.657	28.000	84.734	53.350	2.403.319	**12.244.056**		80	2
Case 1: x86 brand 1	2.157.432	11.906.075	539.000	254.201	112.833	15.527.991	**30.497.533**		1.232	77

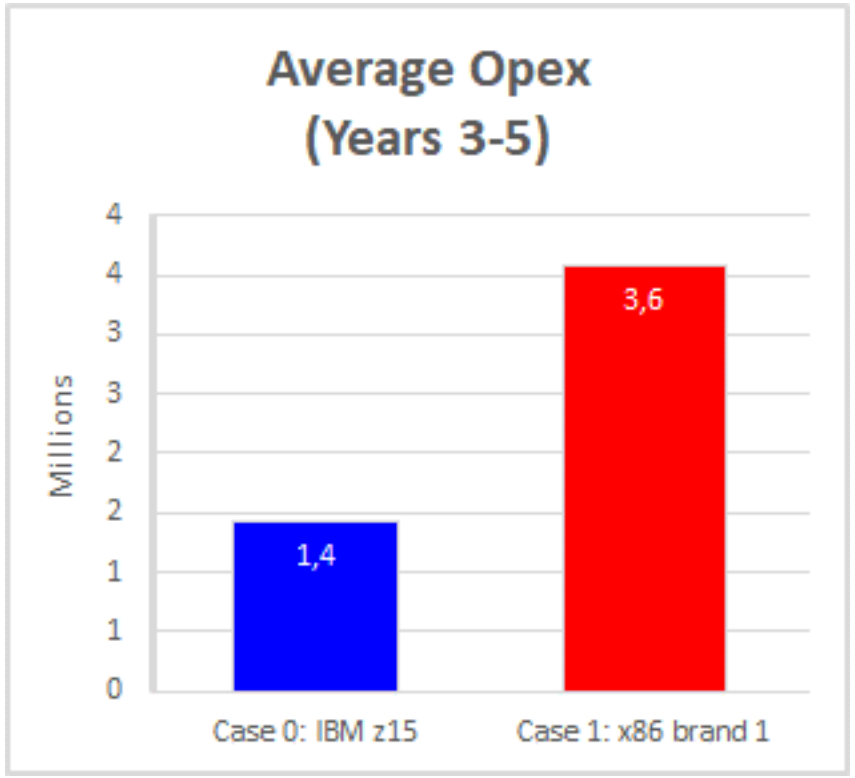


While the x86 Hardware is much cheaper, the other costs are much higher.
The DR requirements increase the differences

*** Data from a Major EMEA bank evaluating the same workload deployed on equivalent environments in terms of performances.**

IT Economics – Linux on z15 vs x86





Beside the savings due to the initial purchase, the Z running costs are also lower, as shown on the Average Opex costs.

*** Data from a Major EMEA bank evaluating the same workload deployed on equivalent environments in terms of performances.**

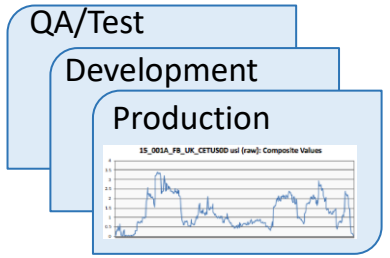
z15 and LinuxONE compared to x86

4 racks
78 servers
1,672 cores

1 rack aka frame
108 cores

z15

Linux workload



34 real customer workloads driving the same throughout and response time

LinuxONE

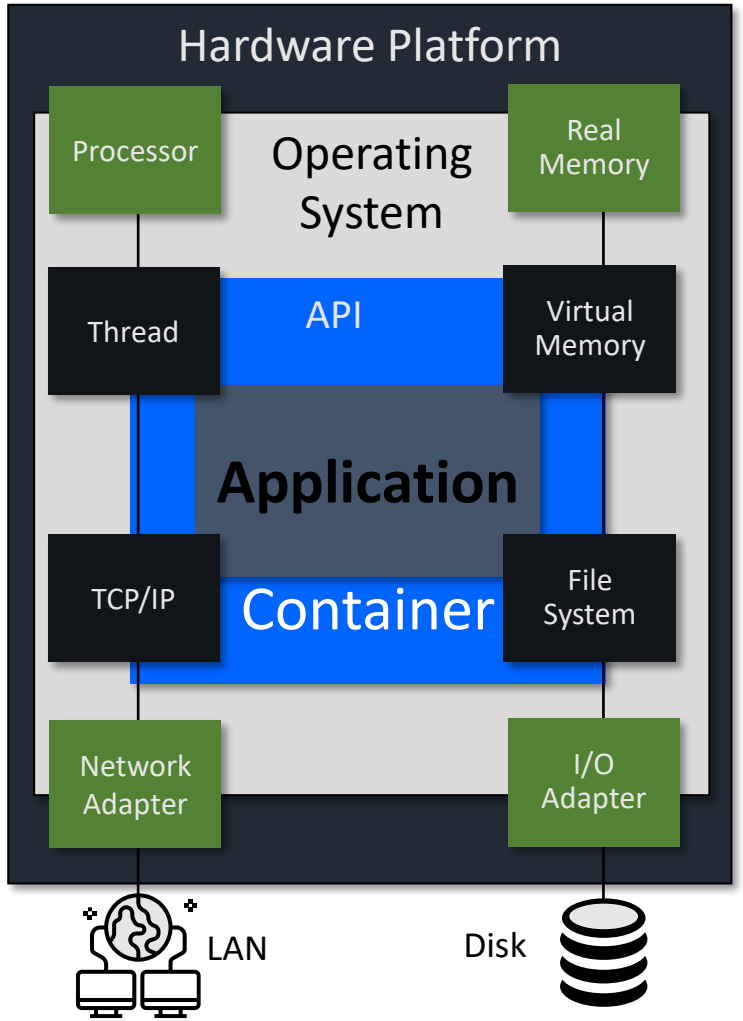
The Z Hardware platform

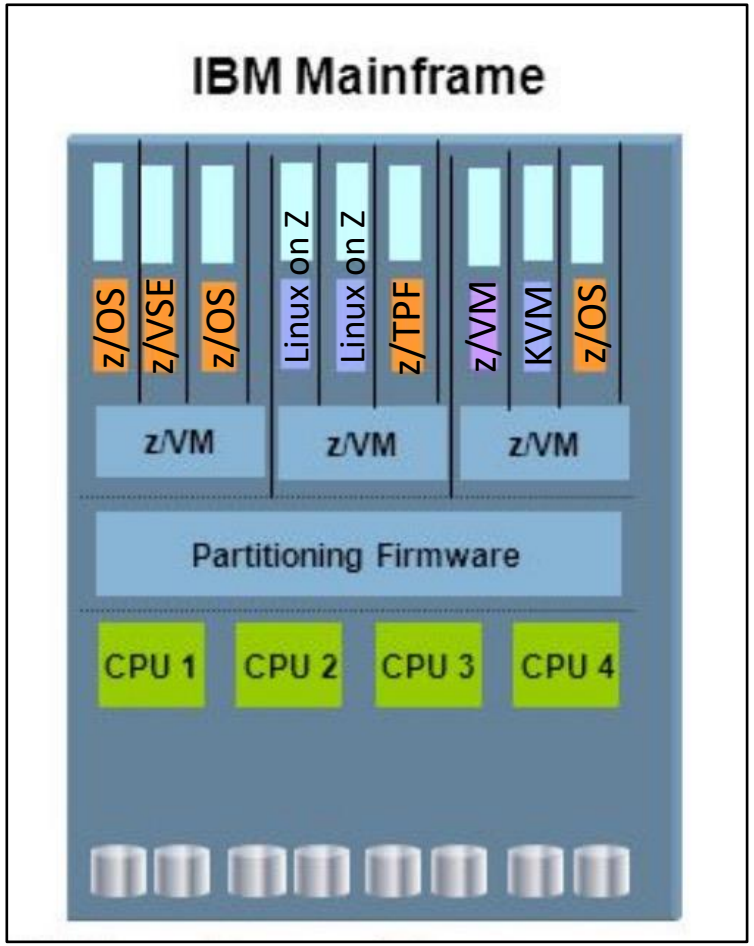
Provider of physical resources

- Processor(s)
- Memory
- Network connections
- Disk connections
- Other I/O connections
- Long term storage
- Printers

IBM Z: Designed for resource virtualization

- Virtualization engine, highly tuned for 50 years, is implemented at Firmware level
- **CPU utilization at 90% routinely**





Z Systems have several types of Processors

Central Processor (CP)

- This processor type is available for normal operating system and application software.

System Assistance Processor (SAP)

- The SAPs execute internal code to provide the I/O subsystem. It manages multiple paths to control units and performs error recovery for temporary errors. Operating systems and applications cannot detect SAPs, and SAPs do not use any "normal" memory.

Integrated Coupling Facility (ICF)

- These processors run only Licensed Internal Code. They are not visible to normal operating systems or applications. A coupling facility is, in effect, a large memory scratch pad used by multiple systems to coordinate work. ICFs must be assigned to LPARs that then become coupling facilities

Integrated Facility for Linux® (IFL)

- This is a processor dedicated to Linux workload.

zAAP

- This is a processor dedicated to execute Java™ code.

zIIP

- The System z Integrated Information Processor (zIIP) is a processor dedicated to execute eligible database workloads.

IFL, zAAP and zIIP engines exist only to control software costs.

Spare

- An uncharacterized PU functions as a "spare." If the system controllers detect a failing CP or SAP, it can be replaced with a spare PU. In most cases this can be done without any system interruption, even for the application running on the failing processor.

- The LinuxONE server has two PU spare per system.
  - If an active PU (Integrated Facility for Linux – IFL, System Assist Processor – SAP, or Integrated Firmware Processor – IFP) fails, the failed PU's characterization is dynamically and transparently reassigned to a spare PU.
  - Transparent sparing for failed processors is supported across the two CPC drawers in the unlikely event that the CPC drawer with the failure does not have spares available.
- This function, managed by the IBM LinuxONE firmware is fully transparent to applications, operating systems, or hypervisors.

- More information available on IBM Redbook SG24-8852 – z15 T02 (8562) technical guide:
  - http://www.redbooks.ibm.com/abstracts/sg248852.html

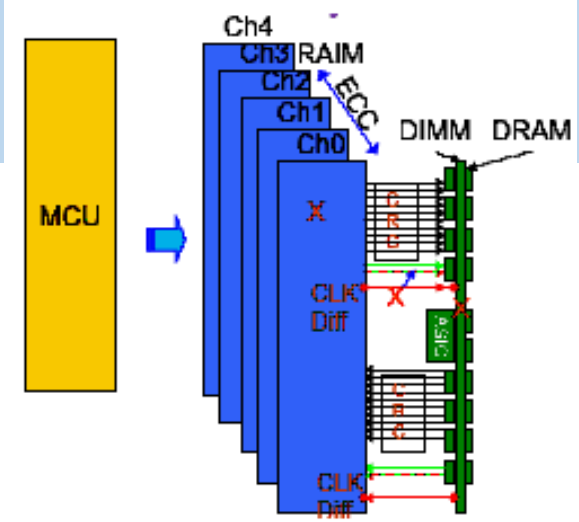
Logical Partitions (LPAR)

- **Processor Resource/System Manager (PR/SM**) allows multiple **logical partitions (LPARs**) to share physical resources such as CPUs, I/O channels and LAN interfaces.
  - **Logical partitions (LPARs)** are, in practice, equivalent to separate mainframes. With minor exceptions z/OS, the operators, and applications cannot detect the difference.
  - Each LPAR can run its own operating system image, independently from the other LPARs, including Linux for Z, z/VM or KVM hypervisor.
  - An LPAR can be added, removed, activated, or deactivated at any time. If the system in one LPAR crashes, there is no effect on the other LPARs.
  - Changing the number of LPARs is not disruptive.

- Processors can be dedicated or shared among LPARs.
- Memory must be dedicated to each individual LPAR.
- PR/SM manages and optimizes allocation and the dispatching of work on the physical topology.

- More information available on IBM Redbook SG24-8851 – z15 T01 (8561) technical guide:
- http://www.redbooks.ibm.com/abstracts/sg248851.html?Open

Memory Sparing



RAIM Technology

- The LinuxONE III servers use the Redundant Array of Independent Memory (RAIM) technology.
  - The RAIM design requires the addition of one memory channel that is dedicated for reliability, availability, and serviceability (RAS).
  - A fifth channel in each Memory Control Unit (MCU) enables memory to be implemented as a RAIM.
  - In case of a Dynamic Random-Access Memory (DRAM), sockets, memory channels, or DIMMs failure, the LinuxONE firmware detects and automatically recovers from failures.
  - This function is fully transparent to applications, operating systems, or hypervisors.
- The five channel RAIM Memory Controller overview is shown on the picture below
- RAIM technology features significant error detection and correction capabilities. Bit, lane, DRAM, DIMM, socket, and complete memory channel failures can be detected and corrected, including many types of multiple failures. Therefore, **RAIM takes 20% of DIMM capacity (a non-RAIM option is not available)**.
- The RAIM design provides the following layers of memory recovery:
  - ECC with 90B/64B Reed Solomon code.
  - DRAM failure, with marking technology in which two DRAMs can be marked and no half sparing is needed. A call for replacement occurs on the third DRAM failure.
  - Lane failure with CRC retry, data-lane sparing, and clock-RAIM with lane sparing.
  - DIMM failure (discrete components and VTT Reg) with CRC retry, data-lane sparing, and clock-RAIM with lane sparing.
  - DIMM controller ASIC failure.
  - Channel failure started RAIM recovery.

Dynamic Memory Relocation

- The LinuxONE server includes 2 CPC drawers.
- It implements dynamic memory reallocation mechanism, which is especially useful during service operations like Enhanced Drawer Availability (EDA) and Concurrent Drawer Replacement (CDR).
- PR/SM controls the reassignment of the content of a specific physical memory array in one CPC drawer to a physical memory array in the other CPC drawer.
- To do accomplish this task, PR/SM uses all the available physical memory in the system.
- This memory includes the memory that is not in use by the system that is available but not purchased by the client, and the planned memory options, if installed.
- More information available on IBM Redbook SG24-8852 – z15 T02 (8562) technical guide:
- http://www.redbooks.ibm.com/abstracts/sg248852.html

The IBM z15 – Mainframe evolution as of today

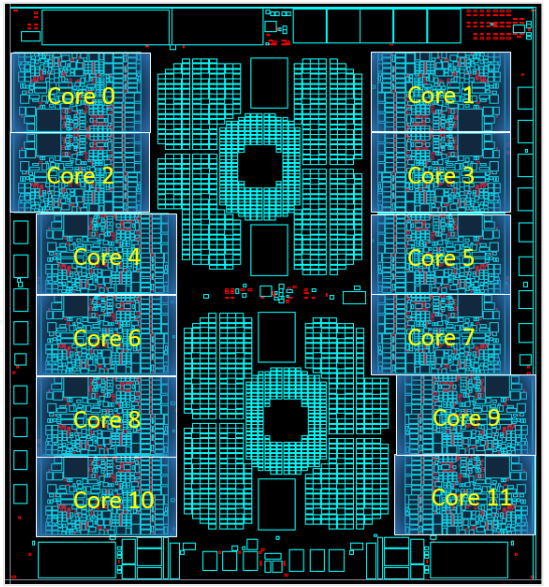
Data Center in a box

- Up to 108 configurable cores on a single rack
- 5.2 GHz
- Up to 40 TB configurable memory
- Back compatible with all systems going back to 1960's
- Leading HPC technologies researches that then flow-down to wider community
- Integrate latest electronic technologies for highest performances

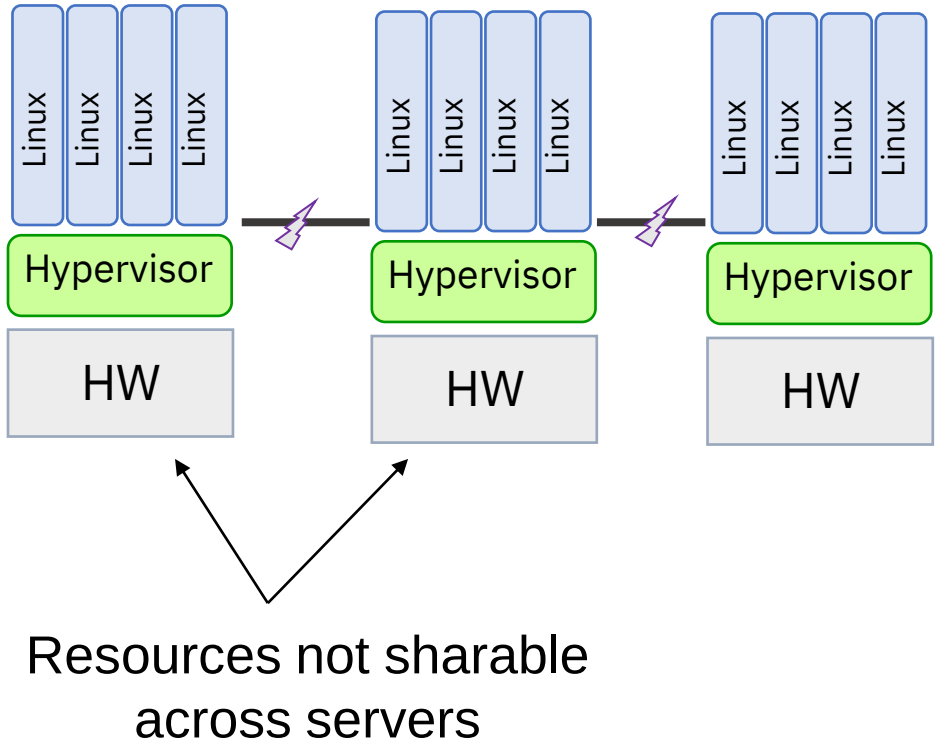
Features like:

- Integrated Accelerator for zEDC
- Crypto Express7S
- System Recovery Boost

… and *many* more



Challenges with the commodity x86 model



- Fixed/limited resources – cores, memory, I/O
- No partitions - have to use separate servers for complete workload isolation
- No capacity on demand
- Less reliable – no spare cores, no RAIM memory
- Low utilizations - stranded resources, provision extra for peaks, growth, reliability
- Network topology can be slow and open to risk
- More time spent in management

What's the right solution to address today's IT challenges?

Is this what your data center looks like?

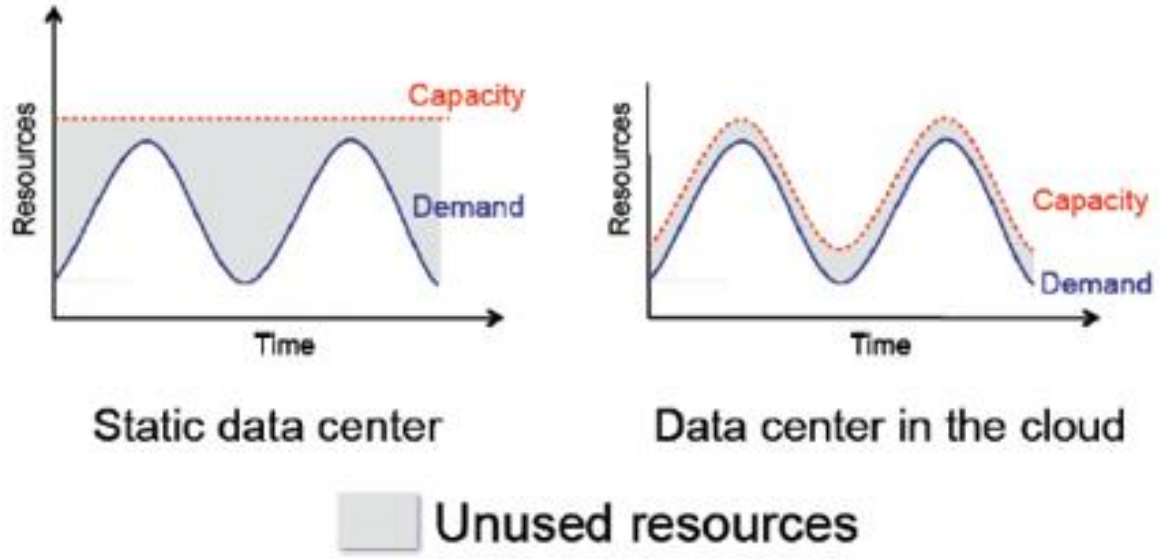
What would change if you would be able to run it on a single, or few, boxes?

- Currently running lots of x86 or scale-out UNIX servers
- Have serious concerns about
  - Downtime
  - Data security
  - Data center floor space and energy usage
  - Growth and scalability
- Strategically committed to Linux, cloud and open source

- Improved efficiency
  - Fewer servers, less networking
  - Fewer software licenses
  - Growth within the box
  - Better utilization of compute resources
- Reduced risk, better security, higher availability
- Reduced costs, reduced staff, simplification

To cope with Peaks typical approach is over-provisioning

- Over-Provisioning: resource provisioning by taking into account peak loads
- Leads to excess capacity and under-utilization
  - Server utilization in traditional data centers is quite low
    - Typically <20%; rarely 30%
- This lead to higher cost than required
- Energy requirements and consumption also increases



Typical server CPU utilization is less than 20%

Detailed performance survey across random selection of several thousand servers hosted in different data centers, located in five continents, across various industries, and over a 2 year period of time:

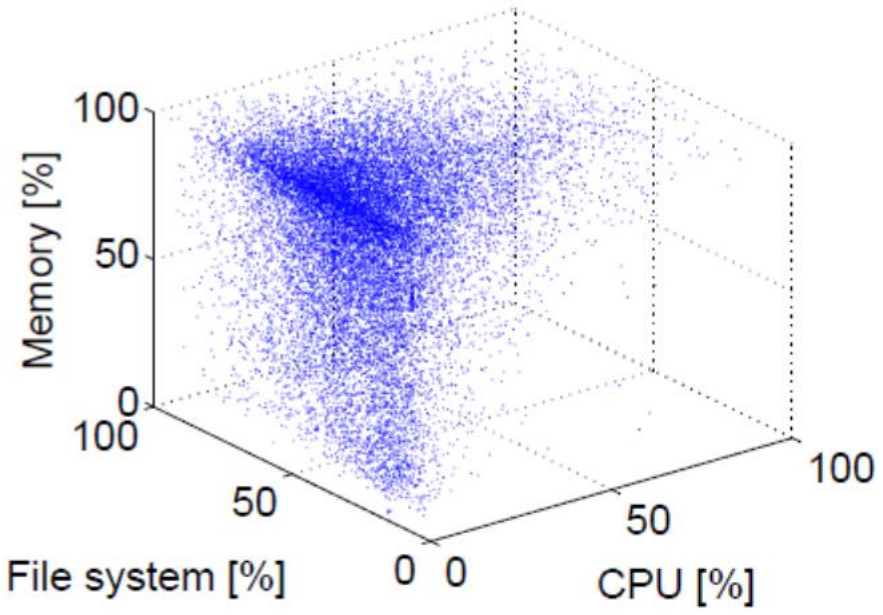


Figure 1: Mean CPU, memory, and file system utilization of all servers ($\overline{U_{i,j}}$).

Table 1: Overview of resource utilization by different categories

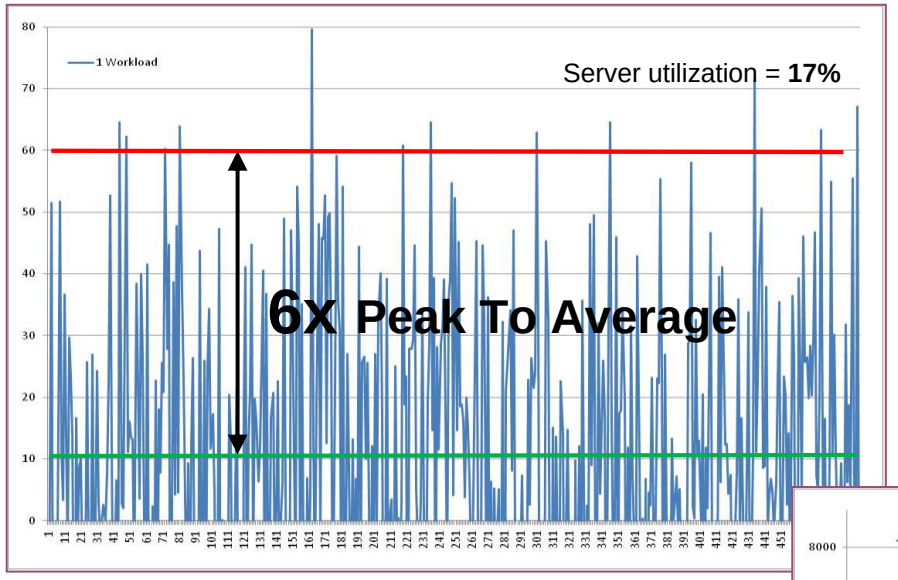
	CPU [%]			Memory [%]			Disk [%]			File system [%]		
All	mean	std	CoV	mean	std	CoV	mean	std	CoV	mean	std	CoV
All	17.76	18.16	1.02	77.93	23.18	0.30	75.28	24.34	0.32	45.17	19.27	0.43
Country												
Country A	24.91	18.30	0.73	80.83	16.09	0.20	81.15	18.37	0.23	55.04	18.35	0.33
Country B	12.89	12.00	0.93	81.73	15.57	0.19	64.10	20.08	0.31	46.33	19.50	0.42
Country C	7.25	9.76	1.35	71.87	25.09	0.35	63.59	27.04	0.43	38.89	18.72	0.48
Country D	14.55	12.93	0.89	84.44	19.02	0.23	70.20	23.98	0.34	48.21	17.95	0.37
Country E	19.28	19.45	1.01	78.57	22.22	0.28	72.33	24.96	0.35	44.13	18.33	0.42
Operating System												
AIX	21.05	19.20	0.91	84.48	16.80	0.20	69.37	23.16	0.33	47.25	18.32	0.39
HP-UX	19.34	16.23	0.84	68.16	20.53	0.30	81.49	17.15	0.21	56.15	16.51	0.29
Linux	6.76	9.73	1.44	80.99	21.29	0.26	85.54	27.17	0.32	36.62	20.24	0.55
Solaris	10.13	11.45	1.13	44.52	23.56	0.53	95.10	12.20	0.13	39.53	20.03	0.51
Type												
Server	12.15	13.87	1.14	71.93	25.96	0.36	78.66	25.77	0.33	44.91	20.47	0.46
LPAR	49.01	20.10	0.85	85.18	16.23	0.19	71.21	22.17	0.31	45.96	17.76	0.39
Solaris Zones	6.20	8.78	1.42	36.65	20.80	0.57	98.42	7.33	0.07	26.63	16.92	0.64

All servers together averaged only 18% utilization

Source: "Data Centers in the Wild: A Large Performance Study", R. Birke, L. Chen, E. Smirni. IBM Research RZ 3820, 2012.

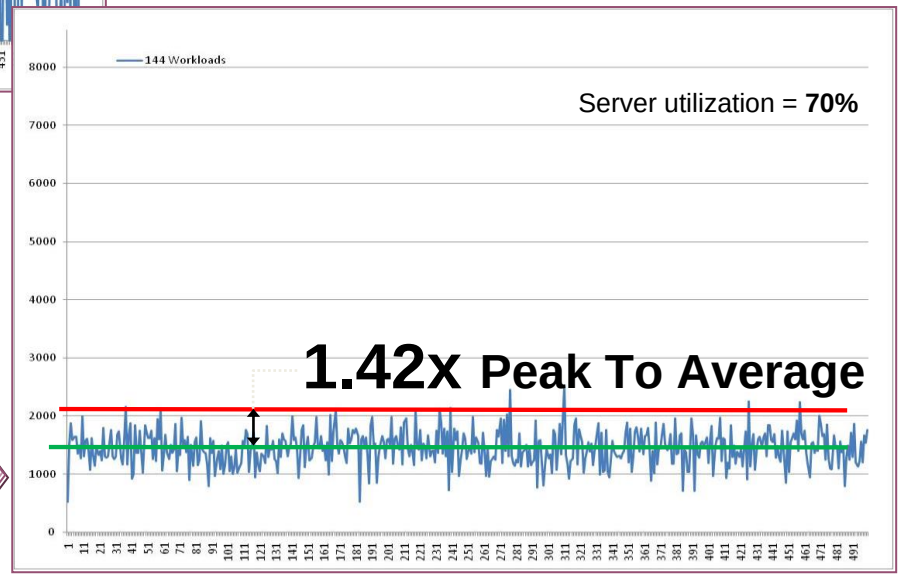
Statistical multiplexing models

Consolidating on a highly virtualized platform drives up CPU utilization



- Consolidating variable workloads on a virtualized server reduces the overall variance (statistical multiplexing)
- Consconsequently, larger servers with capacity to run more workloads can be driven to higher average utilization levels without violating service level agreements

1 variable workload:
Machine capacity (red) =
6x average demand (green)

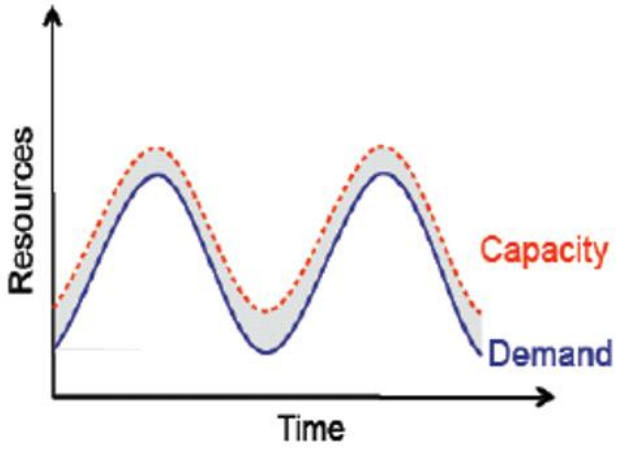


144 variable workloads:
Machine capacity (red) =
1.42x average demand (green)

Moving to cloud to leverage Elasticity

Elasticity is the degree to which a **system is able to adapt** to workload changes by provisioning and de-provisioning resources in an **autonomic** manner, such that at each point in time the **available resources *match* the current demand as closely as possible**

- N.R. Herbst et al., "Ready for Rain? A view from SPEC Research on the Future of Cloud Metrics", Tech Rep. SPEC-RG-2016-01, 2016



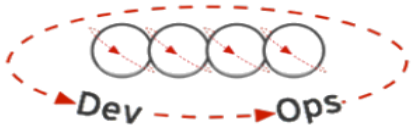
Data center in the cloud

Unused resources

Digital Transformation is changing the way business is done

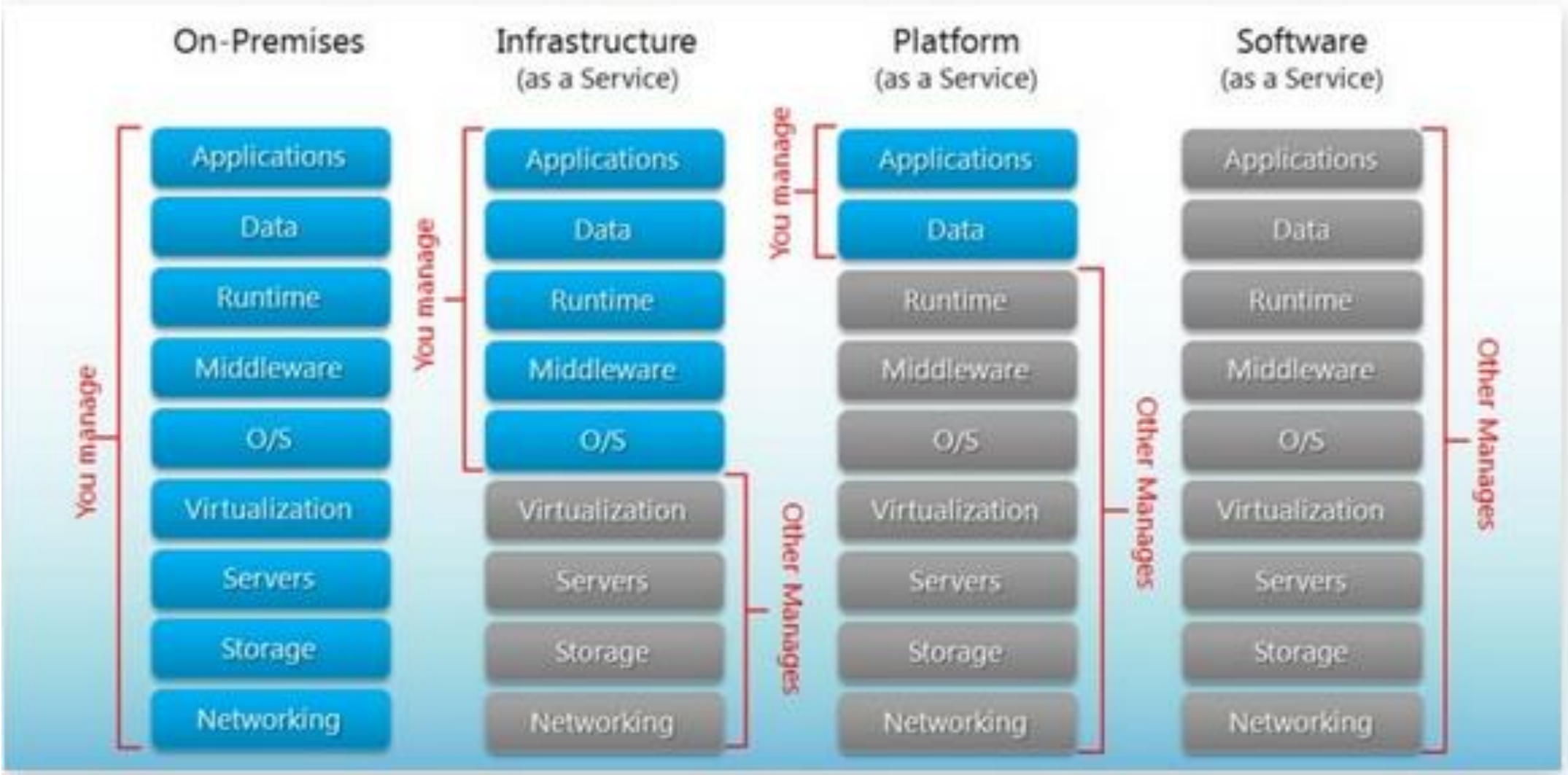
IT Must Evolve to Stay Ahead of Demands

Development Process	Application Architecture	Deployment & Packaging	Application Infrastructure
Waterfall	Monolithic	Physical Servers	Datacenter
Agile	N-Tier	Virtual Servers	Hosted
DevOps	**Microservices**	**Containers**	**Cloud**



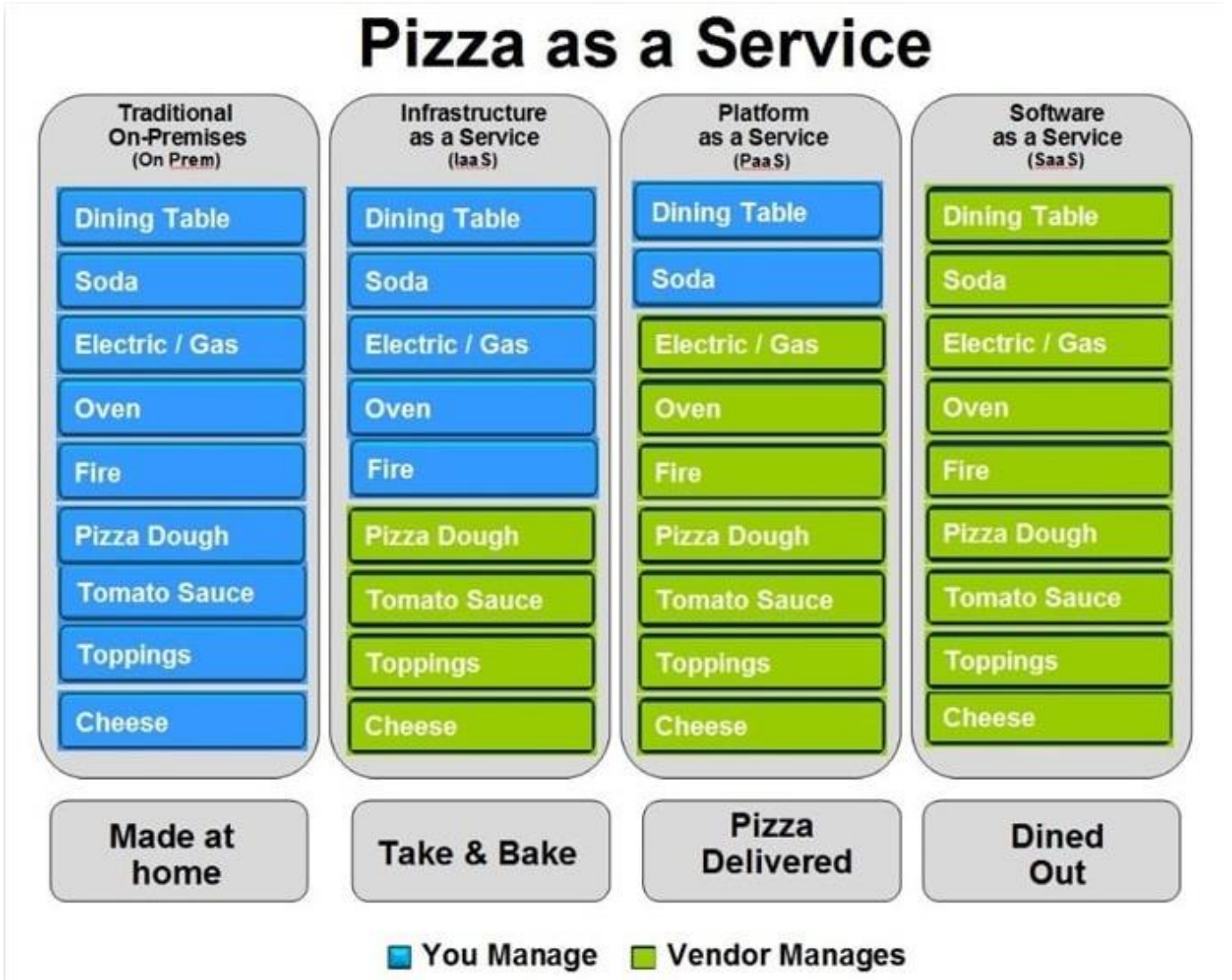
While supporting legacy Business Critical application

Cloud approaches: IaaS, PaaS and SaaS



- Source: https://www.bigcommerce.com/blog/saas-vs-paas-vs-iaas/#the-key-differences-between-on-premise-saas-paas-iaas

Understanding the Cloud approaches: Pizza as a Service

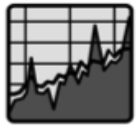


- Source: https://www.bigcommerce.com/blog/saas-vs-paas-vs-iaas/#the-key-differences-between-on-premise-saas-paas-iaas

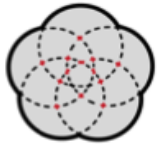
The market is entering a new phase in cloud and digital

Phase 1	*Phase 2*
Consumer-driven innovation	Enterprise-driven innovation
Digital/AI experimentation	Digital/AI embedded in the business at scale
“User applications”	“Mission critical” workloads
Public cloud	Hybrid cloud

The modern vision of the Cloud Strategy



Optimize the IT you have



Integrate apps, data, & processes

Add & manage cloud infrastructure

Evolve existing Build more modern applications

Automate & manage IT

People Skill

1

Hybrid

Enable enterprises across traditional, private and public environments

2

Multi-cloud

To avoid vendor Lock-in

3

Open

Build capabilities that are open by design, enabling client flexibility and reducing vendor "lock in"

4

Secure

Provide reliability and continuous security for the client's environment

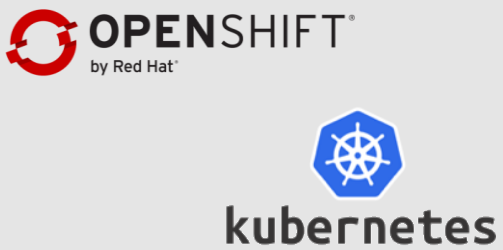
5

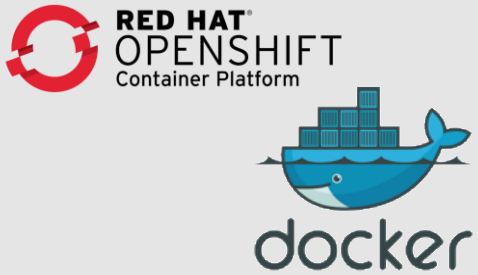
Management

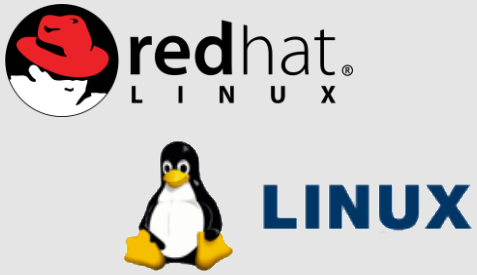
Consistent service level, support, logging, management and delivery across complete cloud environment

Affordable way to access Enterprise grade IT Infrastructure, such as Z, as well as new coming technologies such as Quantum Computing

The importance of open







Container Orchestration

- Manages scalable deployment and lifecycle management
- Large open source community & rapid enterprise adoption
- De facto choice for new applications

Containers

- Enables portability across hardware platforms and clouds
- Accelerating enterprise adoption
- Integrates Development and IT
- Open source and strong community

Operating Systems

- #1 Enterprise OS across cloud providers
- Open source with strong community
- Broad hardware vendor support

Client Value

- Reduce vendor lock-in
- Portability
- Developer reach
- Ecosystem
- Network effect

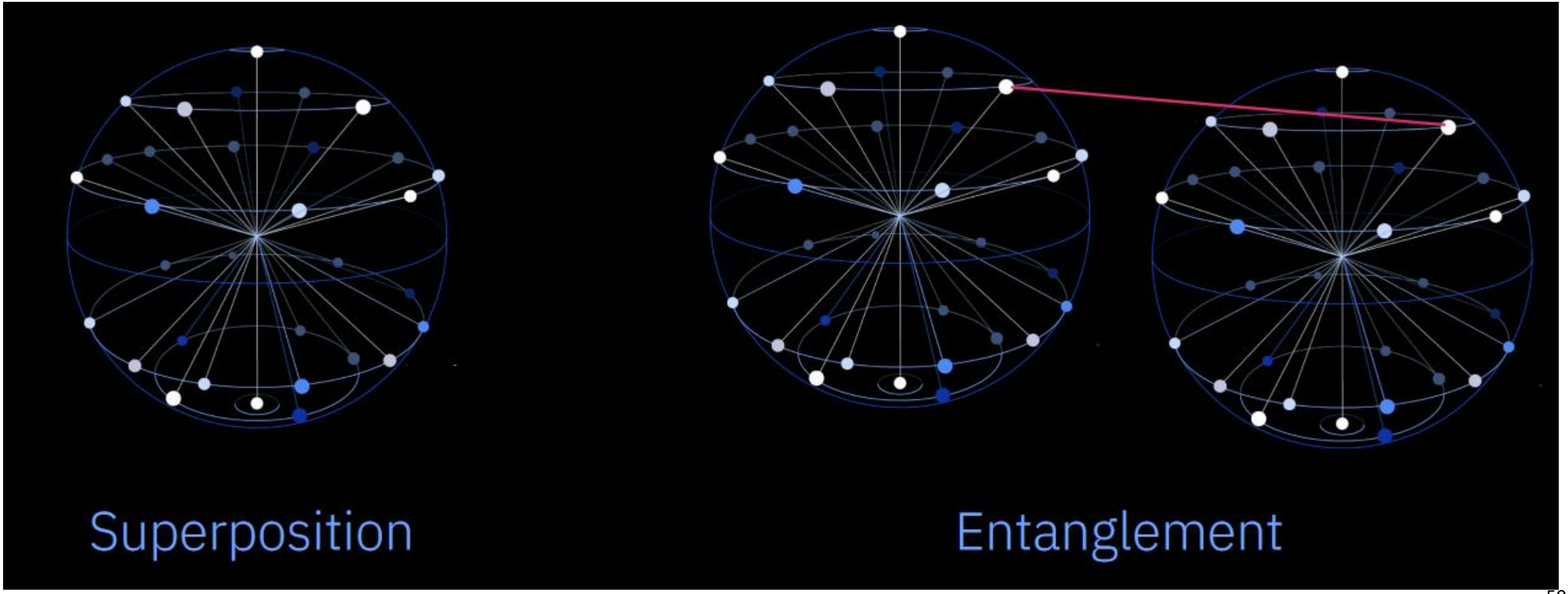
Conclusion and message to bring home

- NFRs are many and very stringent
  - Laws, Rules and **compliance requirements impose many NFRs that have high cost implications** (ex. 24x7 avail.)
  - Security is a NFR that tend to be approached as unexpected event. **We must have a Security approach** to reduce likelihood and cost impacts that could be huge
    - Many technical **solutions are complex and often not appreciated until there are issues**
- Enterprise IT requires high computation needs
  - Common knowledge of x86 technology is that this platform could cover all needs and that Enterprise systems, like Mainframes, are legacy environments now obsolete
    - High level introduction to IBM **Z major features**
  - **Virtualization helps increase system utilization**
    - The more a system is utilized the less will be the TCO
- IT Economics assessment principles: **TCO is very different than TCA**
  - A 5 year TCO can present surprising messages on what is more expensive
    - The Z platform is **cost-effectiveness for Enterprise grade IT Solutions** and to satisfy NFRs
- Cloud Computing is changing the approach to IT Computing
  - Yet **Cloud Computing** doesn't mean to move from On-Premise to just Public Cloud
  - Enterprises have **hybrid cloud** strategy made of multi-public cloud as well as private cloud adoption
  - The cloud approach can help **optimize resource usage**, by workload deployment on production environment made of many containers as well as other resources that optimize performances based on workload
  - The Z platform is key element in hybrid cloud workload placement and **the cloud bring it to wider adoption beside Enterprises**
  - **Quantum Computing** is coming and will be available via Cloud

THANK YOU

What is Quantum Computing?

- Universal quantum computers leverage quantum mechanical properties of superposition and entanglement to create states that scale exponentially with number of **qubits, or quantum bits**.



Quantum Properties

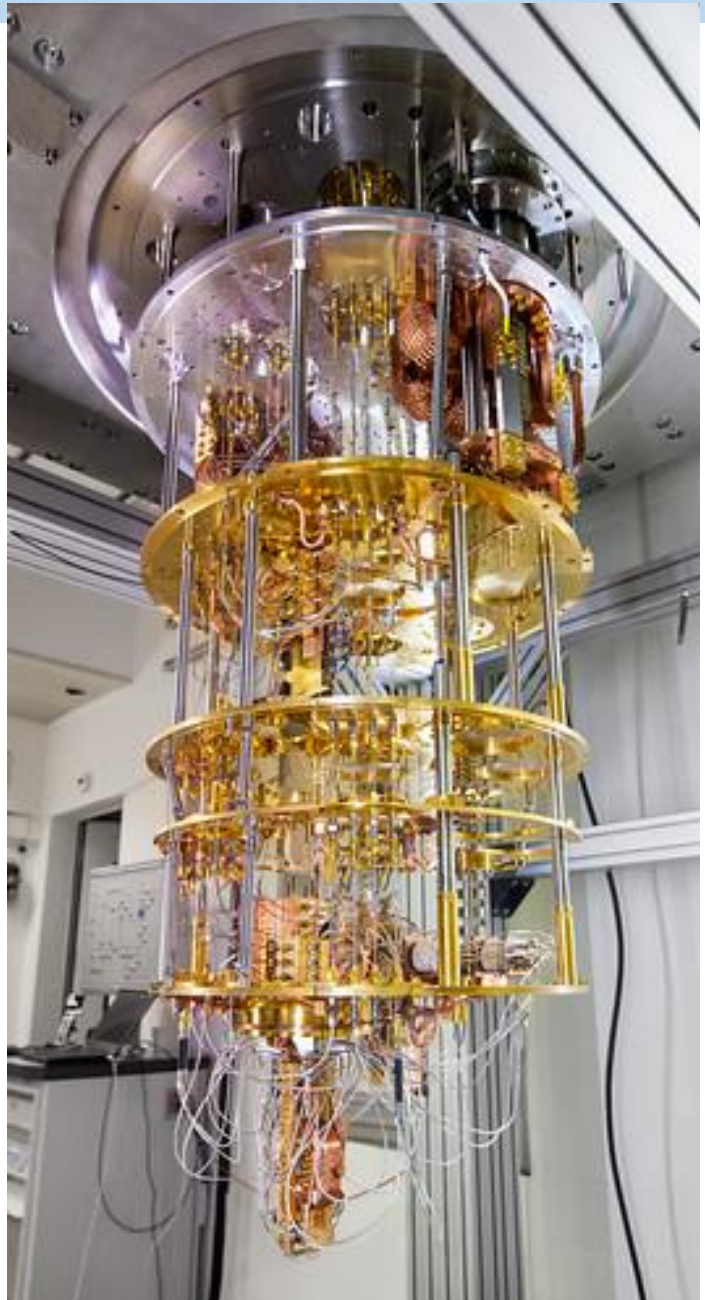
Quantum Superposition

- Quantum superposition is a wavelike property of a quantum system where the system exists in several quantum states at the same time.
- Particles exist across all the possible states with simultaneous varying probabilities.
  - However, once a measurement of a particle is determined, the superposition is lost, and the particle is in one known state.
- Because qubit states are in quantum superposition of 1 and 0, the probability of measuring 0 or 1 for a qubit is neither 1.0 nor 0.0, but a probability in between.
  - Comparatively, the value of a classical bit can only be 1 or 0.

Quantum Entanglement

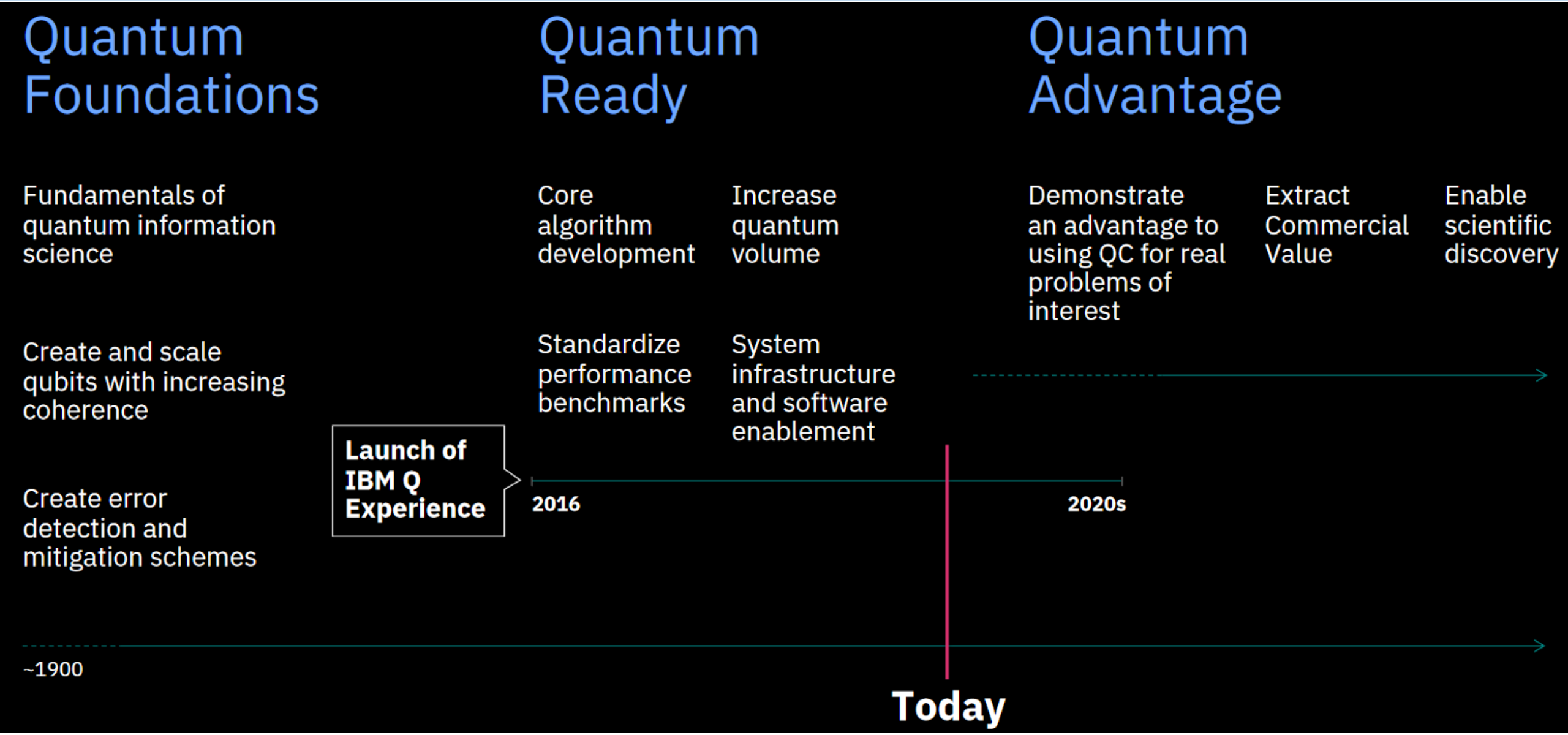
- Quantum entanglement is the quantum physical phenomenon that occurs when a system of particles is generated, interact, or share spatial proximity in ways such that the quantum state of each particle must be described with reference to each other even when separated by a large distance.

A quick look to IBM Q



- https://www.youtube.com/watch?v=o-FyH2A7Ed0

Where are we on the road to Quantum Advantage?



Potential Use Cases for Quantum Computing

	Chemicals and Petroleum	Distribution and Logistics	Financial Services	Health Care and Life Sciences	Manufacturing
● Chemical Simulation	Chemical product design Surfactants, Catalysts			Drug Discovery Protein Structure Predictions	Materials Discovery Quantum Chemistry
■ Scenario Simulation		Disruption Management	Derivatives Pricing Investment Risk Analysis	Disease Risk Predictions	
▲ Optimization	Feedstock To Product Oil Shipping / Trucking Refining Processes	Distribution Supply Chain Network Optimization Vehicle Routing	Portfolio Management Transaction Settlement	Medical/Drug Supply Chain	Fabrication Optimization Manufacturing Supply Chain Process Planning
◆ AI/ML	Drilling Locations Seismic imaging	Consumer Offer Recommender Freight Forecasting Irregular Behaviors (ops)	Finance Offer Recommender Credit/Asset Scoring Irregular Behaviors (fraud)	Accelerated Diagnosis Genomic Analysis Clinical Trial Enhancements	Quality Control Structural Design & Fluid Dynamics

Quantum Computing will be next to traditional computing and will not replace it

Three main research streams on Quantum Computation

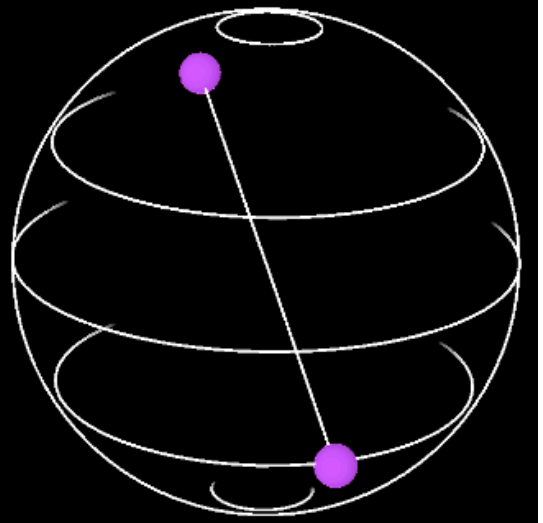
Quantum Computers

IBM **Q** Systems

Quantum Languages

IBM **Q** Experience

Quantum Applications



QISIT AQUA

Access Quantum Computing tools and resources available to researcher and developers from: https://www.ibm.com/quantum-computing/