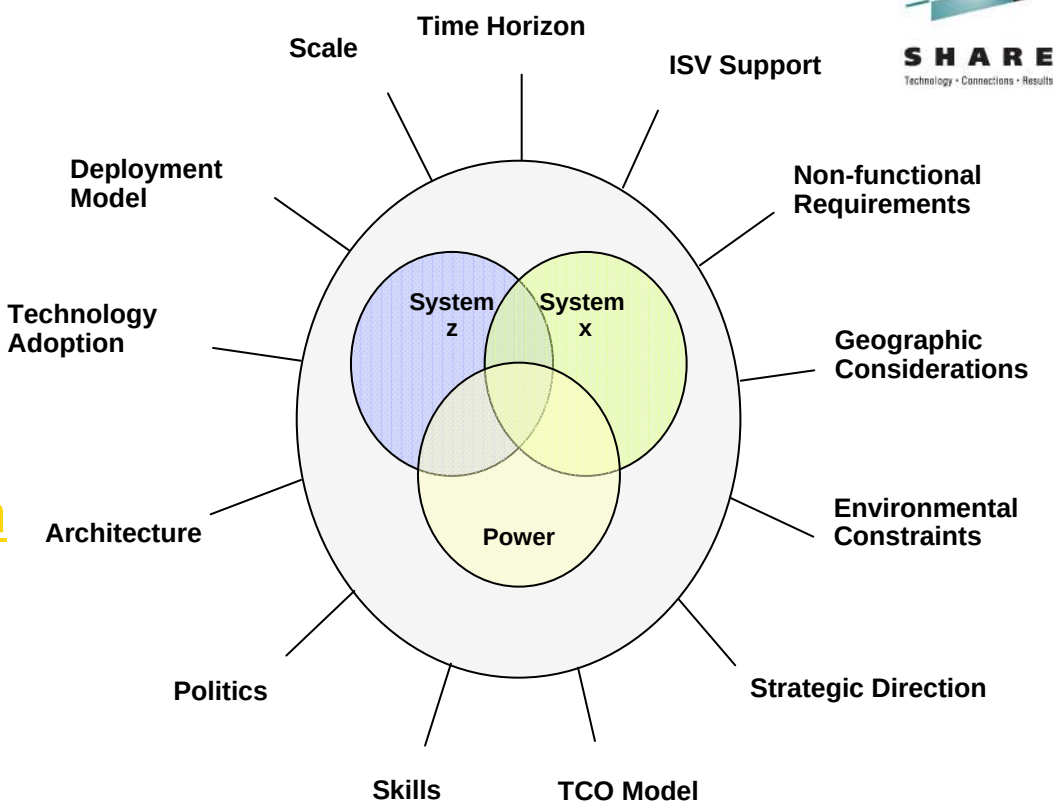
Fit For Purpose

Joe Temple
IBM
jliitemp@us.ibm.com

8/9/2011
Session 09772



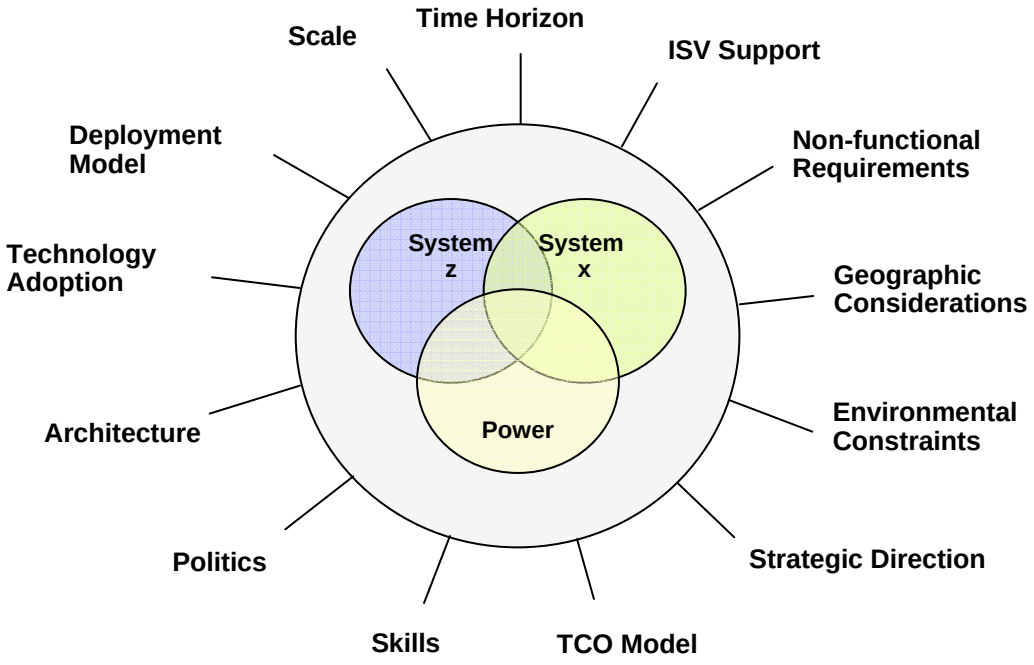
Key IBM Initiatives

- Smarter Planet
- Smart Computing
- Big Data
- Workload Optimization
- Cloud
- *Fit for Purpose*

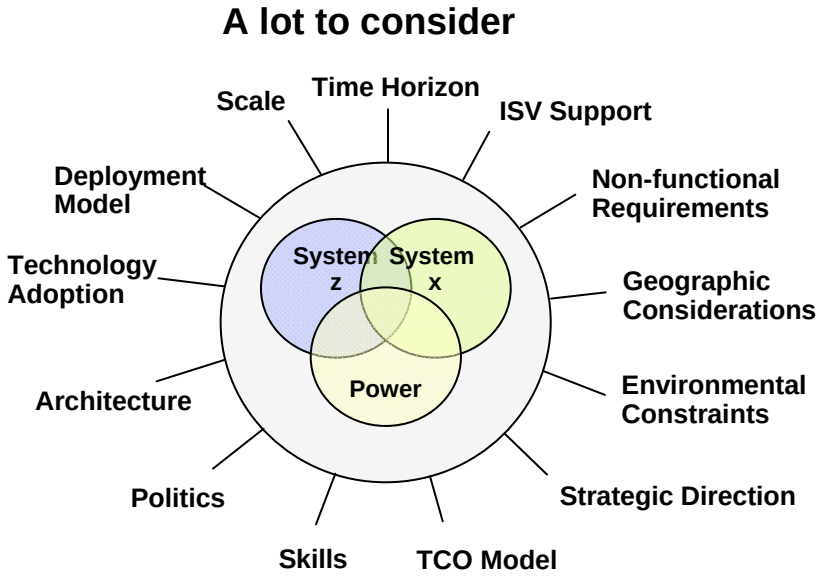
What is F4P?

Fit for Purpose is a client centric thought process that when applied yields rational platform choices which are in line with the client's requirements and local conditions.

It is based on the fundamental principles that "one size does not fit all" and that "local factors matter".

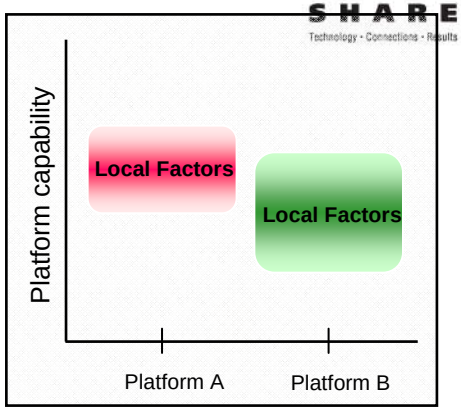


In a Nutshell 1



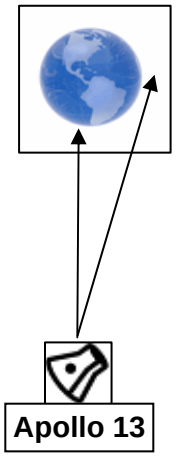
Local Factors Matter

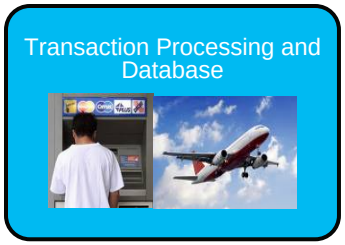
- Skills
- Technology adoption
- Management
- Volume of servers
- Organizational

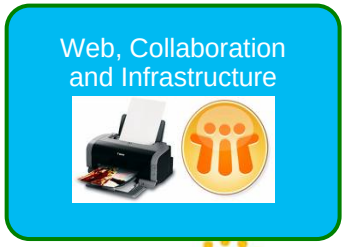


Scale Matters

- Changes people dynamics
- Increases handoffs
- Affects testing, patching, etc





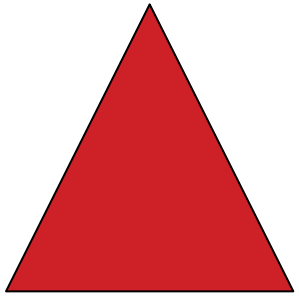


Workloads Matter

SHARE in Orlando 2011

In a Nutshell 2

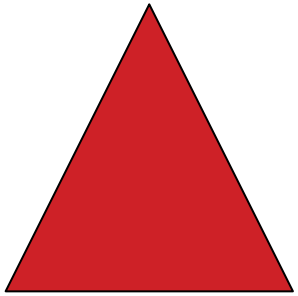
Service Level



Throughput Efficiency

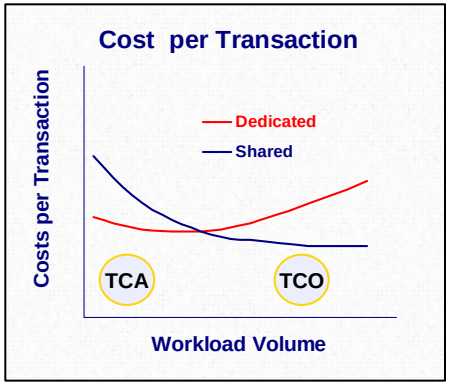
There are operational trade offs

Serial Fitness

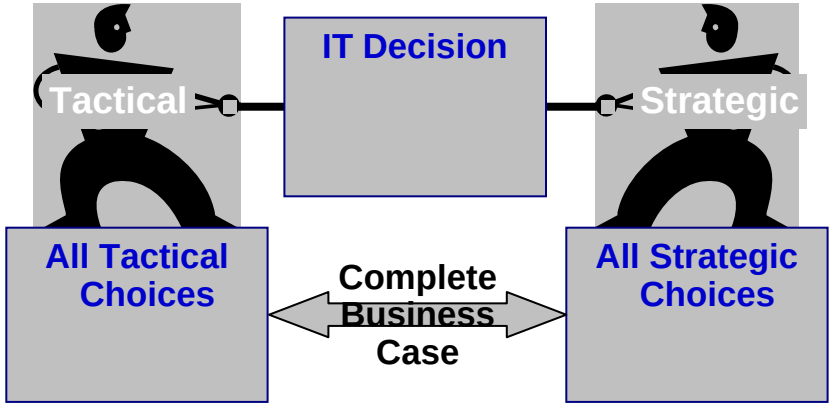


Parallel Fitness Data Fitness

There are server design trade offs



Beware of the hidden cost of sub optimization
Chargeback models often distort the selection process



Strategic and tactical balance matters

SHARE in Orlando 2011

Workload Optimization: F4P is not the same thing

Workload Optimization – One size does not fit all

Analytics and High Performance

Transaction Processing and Database

IBM has used IDC market segments to differentiate workloads

Web, Collaboration and Infrastructure

Business Applications

Fit for purpose views the same segments differently

An SAP solution will exhibit any or all of these types

Mixed Workload – Type 1

- Scales up
- Updates to shared data and work queues
- Complex virtualization
- Business Intelligence with heavy data sharing

Parallel Data Structures – Type 3

- Scales well on clusters
- XML parsing
- Buisness intelligence with little data sharing
- HPC applications

Application Function *Data Structure* ***Usage Pattern*** ***SLA*** ***Integration*** ***Scale***

Highly Threaded – Type 2

- Scales well on large SMP
- Web application servers
- Single instance of an ERP system
- Some partitioned databases

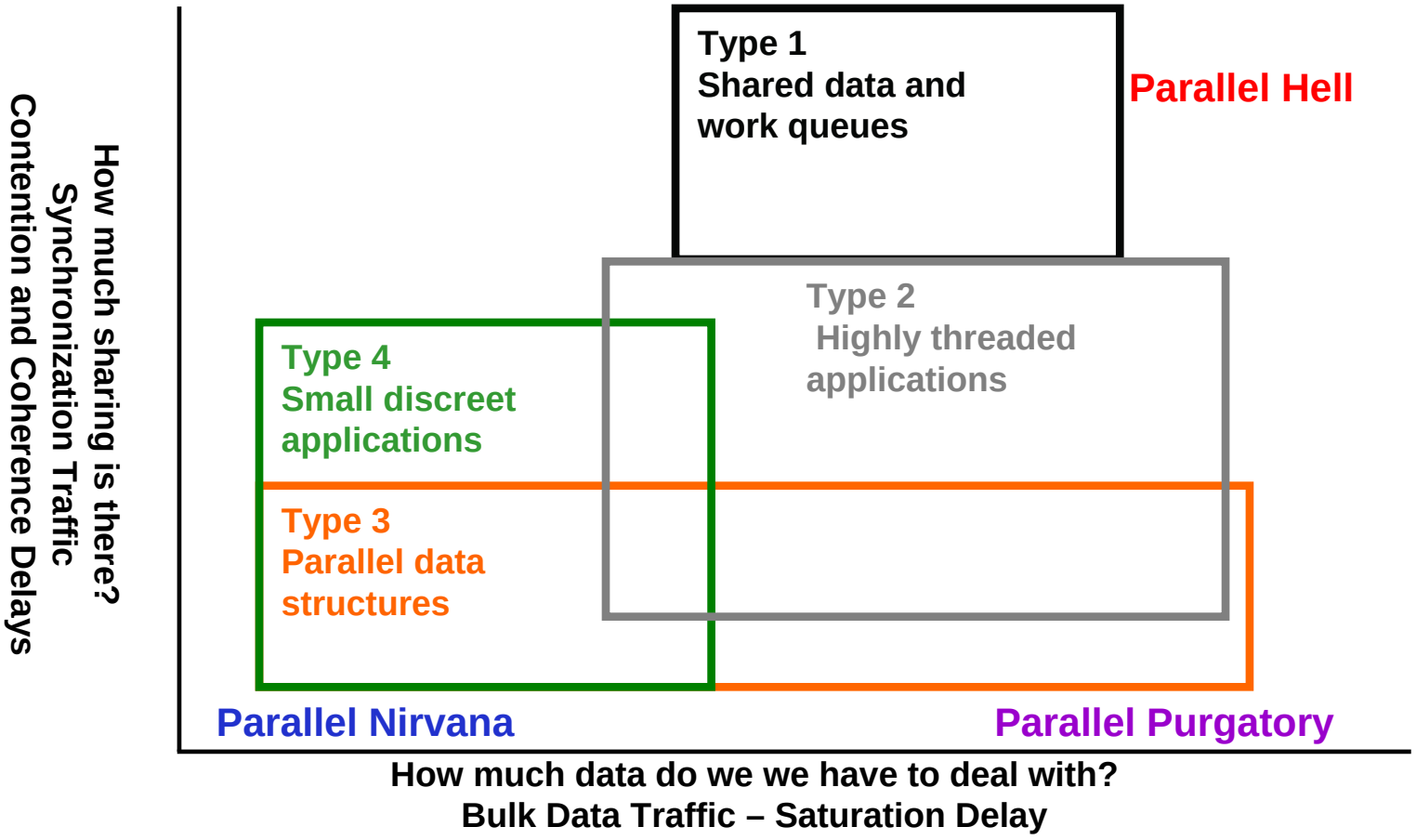
Small Discrete – Type 4

- Limited scaling needs
- HTTP servers
- File and print
- FTP servers
- Small end user apps

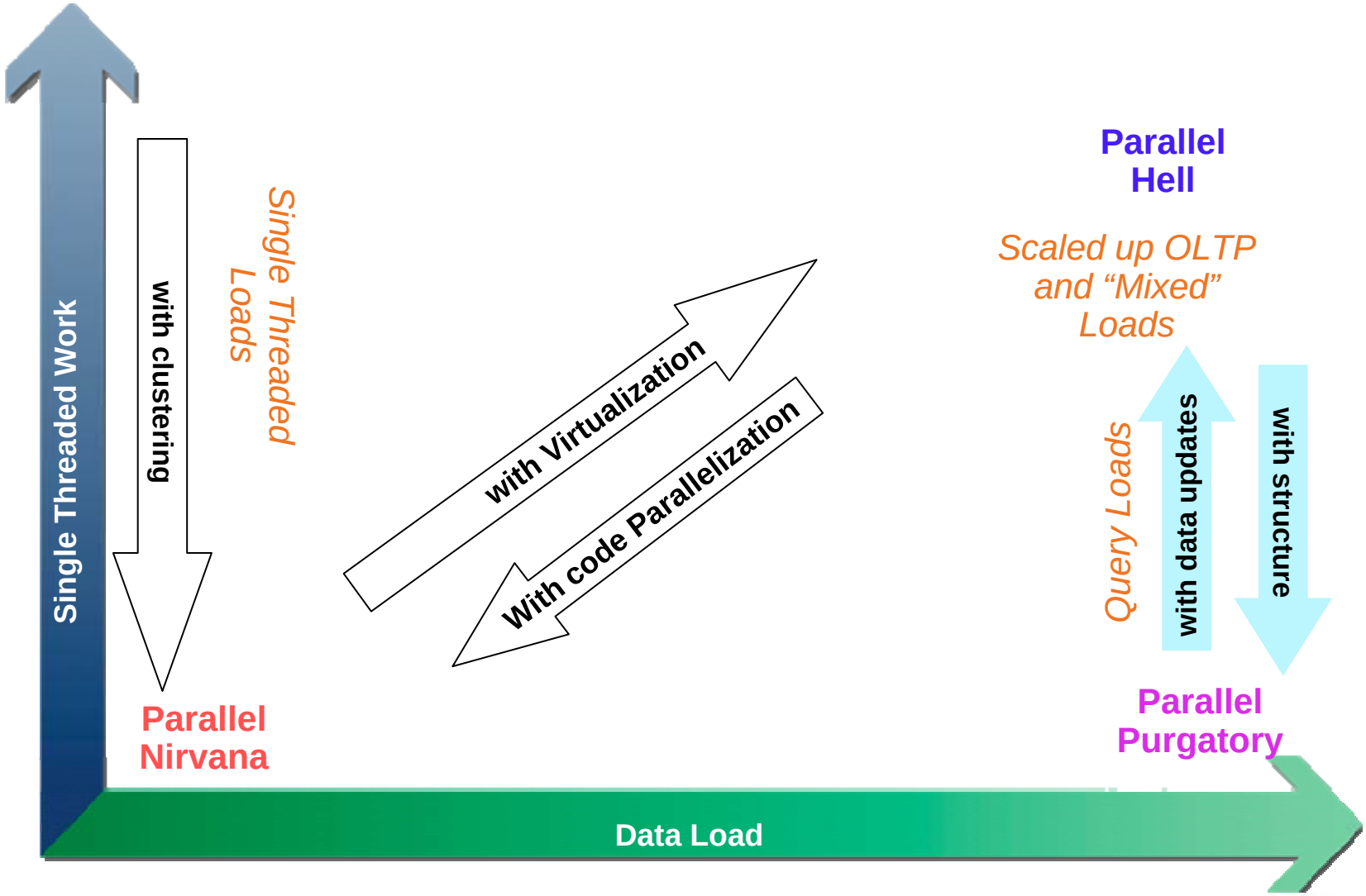
Black are design factors

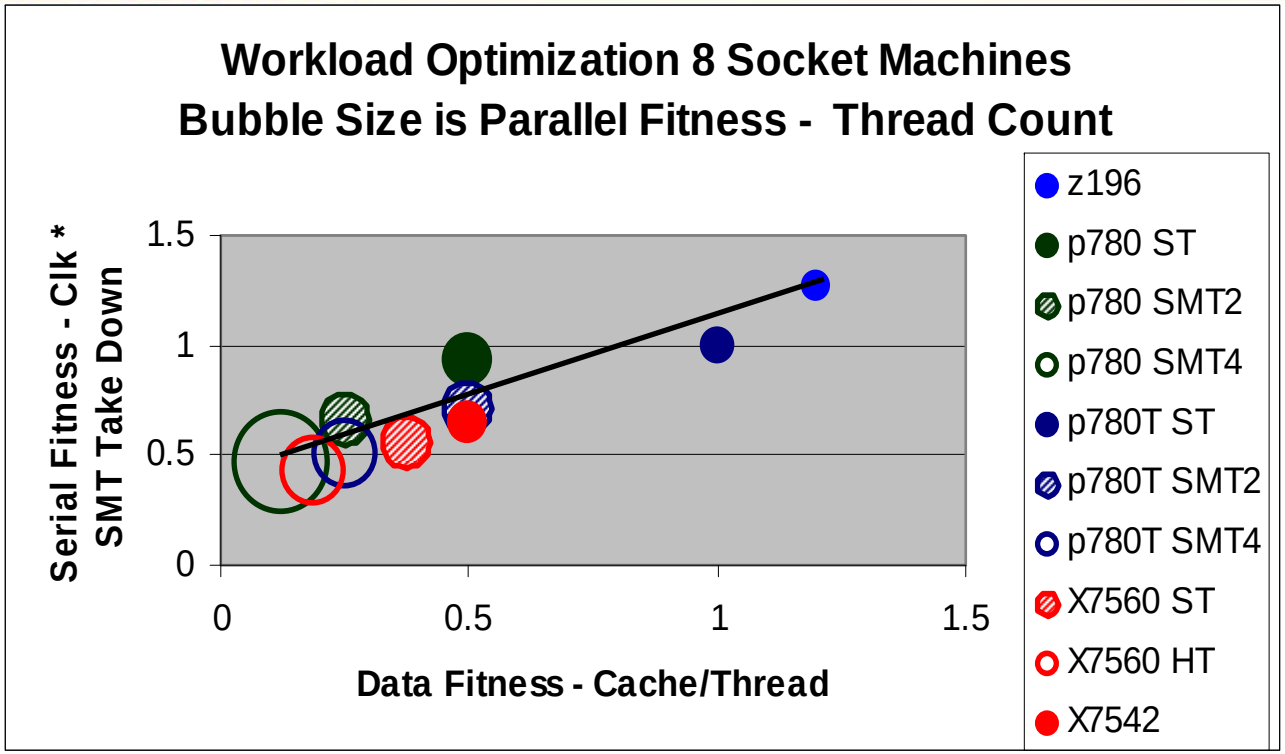
Blue are local factors

Workload Types and Pfister's Paradigm



Pfisters Paradigm:
One size does not fit all





You can put any machine on this chart
All you need to know is:
Clock Speed
Top Level Cache Size and Sharing Scope
Core Count and Thread Count per Core
Throughput Claim for Multithreaded Cores

The SMT Take down is 1/SMT throughput multiplier. This is an application of "Little's Law"

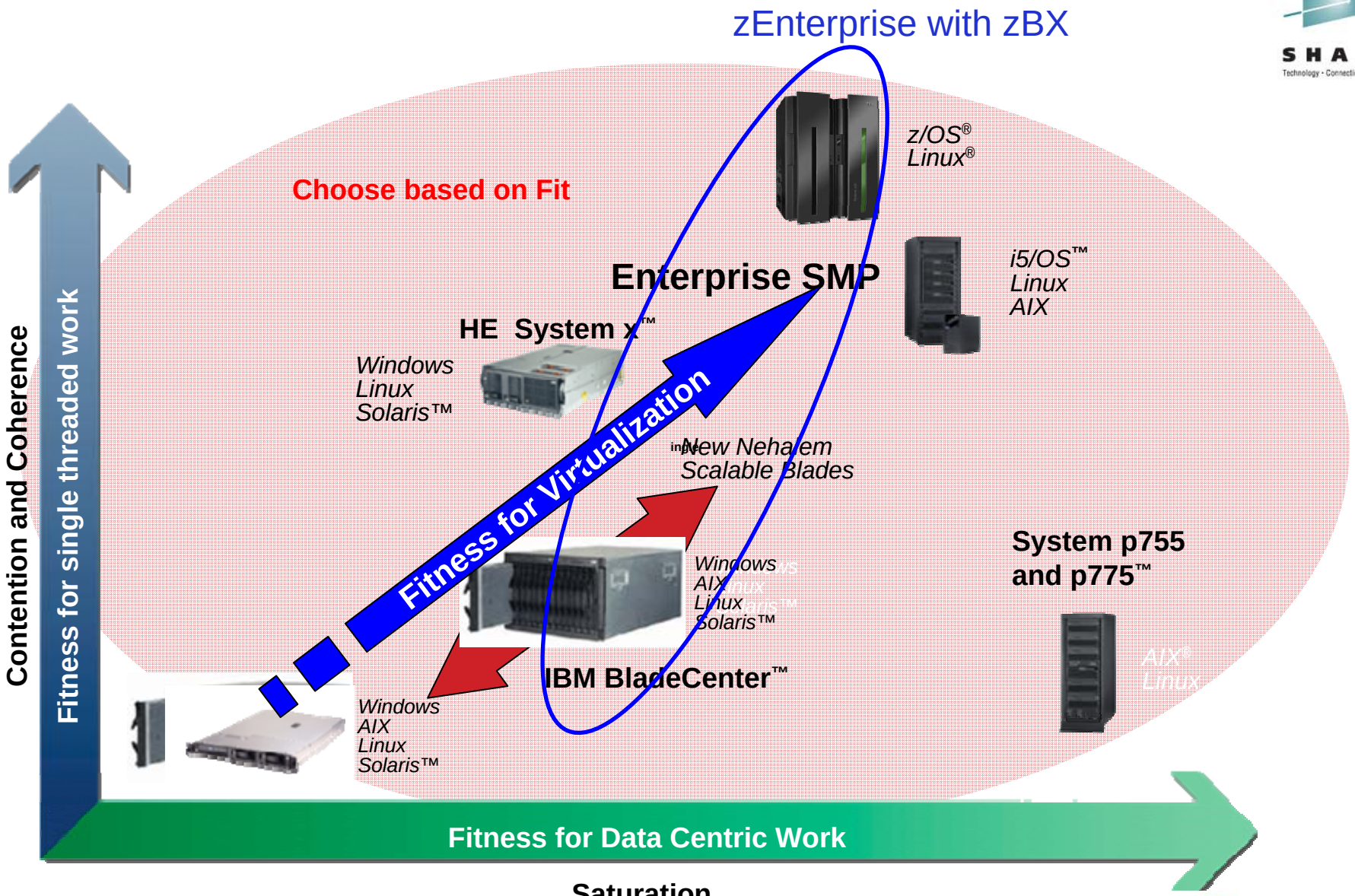
The trend line comes out of the page on the lower left.

There is a clear trade off at work here:

To have more concurrent threads you must give up thread speed and cache/thread

Note: Each brand will define capacity in an advantageous way which is why fundamentals work better than benchmark derived metrics.

IBM Systems are fit for purpose and cover key legacies

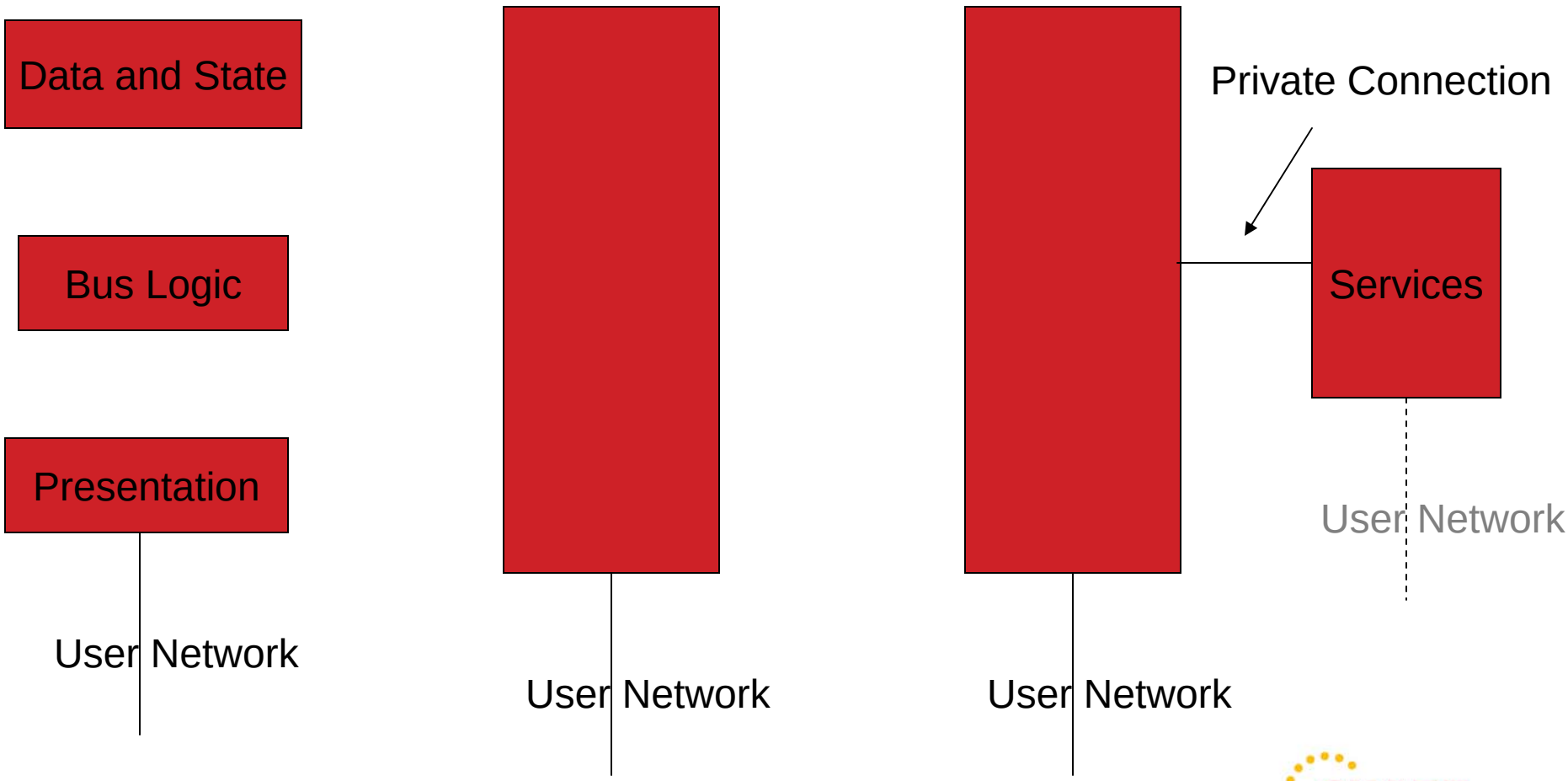


“Hybrid Computing” turns “Client Server” on its ear

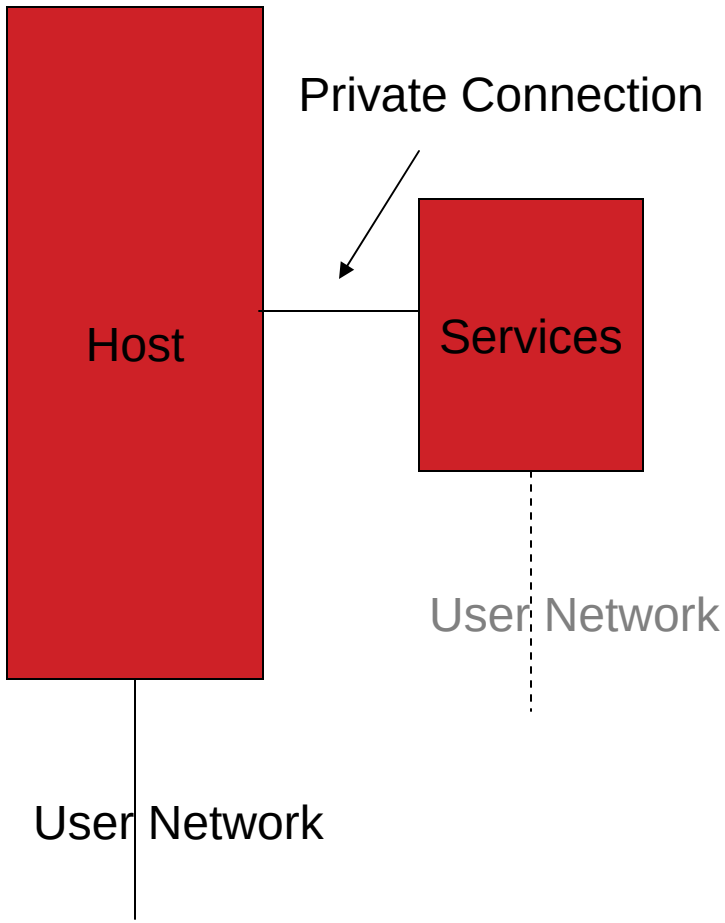
Client Server Tiers

Centralized Sharing

Hybrid Computing



Hybrid Computing



Hybrid APIs available today

Open CL Language
- Math accelerator API

WebSphere Compute Grid
- JAVA API

CoZ Launcher
- Batch to secure shell API

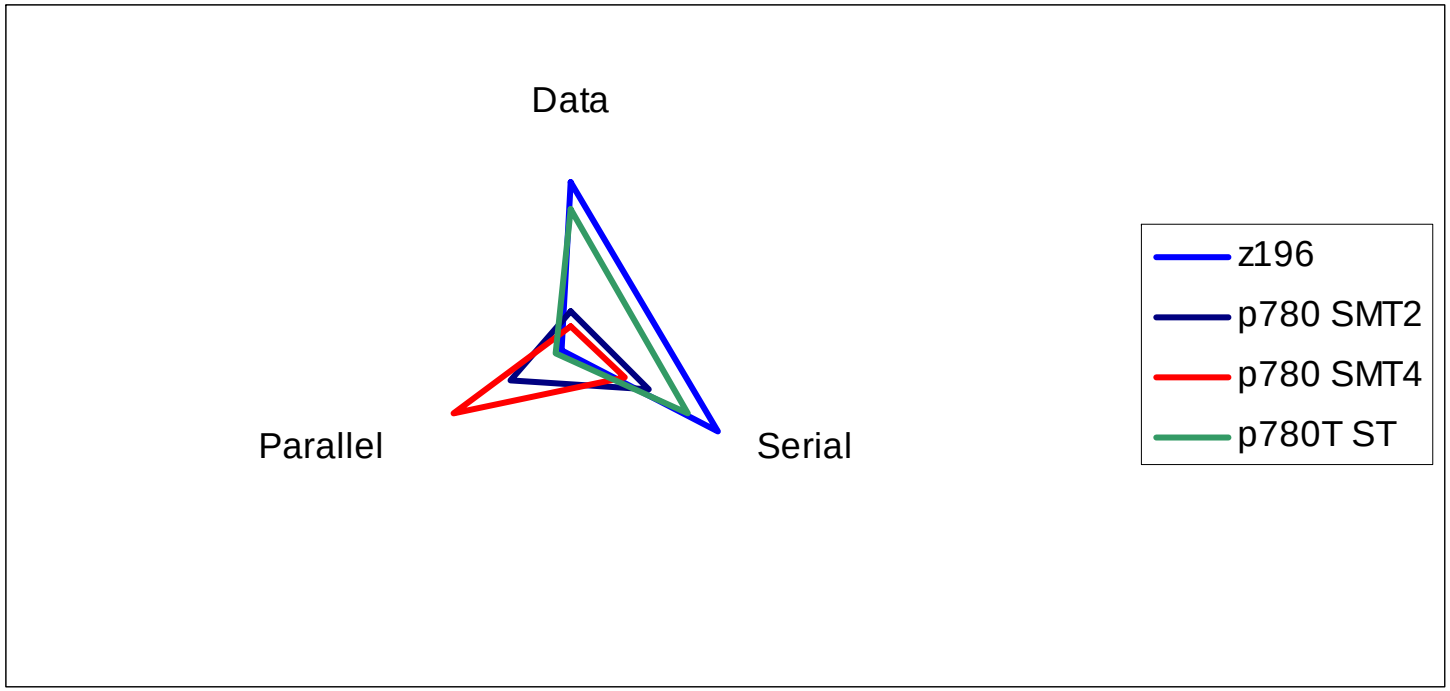
Cloud is not about deployment model or server type.

- Cloud is simply a layer of software
  - Fills gaps from underlying platform in workload management, provisioning, etc.
  - Provides for self service by clients
  - Provides a chargeback mechanism by which clients rent services
    - IAAS Infrastructure as a Service
    - SAA Software as a Service
    - AAAS Application as a Service
    - DAAS Data as a Service
    - BPAAS Business Process as a Service

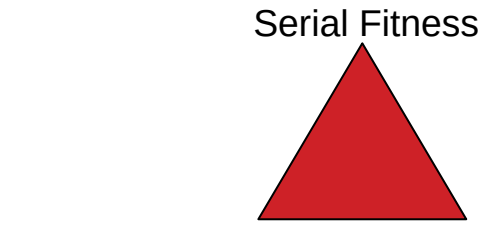
Clouds are platforms

- When designing them there are fit for purpose decisions in choosing the underlying infrastructure
- When employing them there are fit for purpose decisions about using them

Another way to look at the data



This leads to the notion of a “trade off triangle”



Sign in my office:

It’s about the Client, stupid

Local factors not only matter, they rule the day

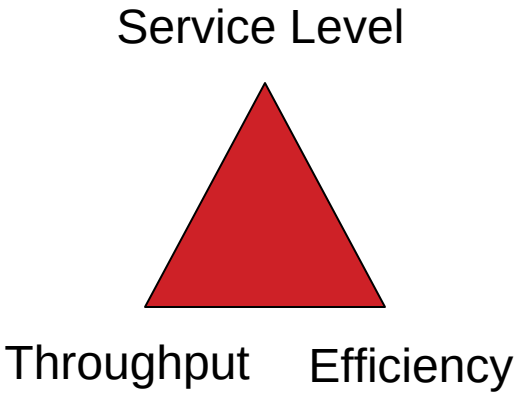
Server positioning is not enough

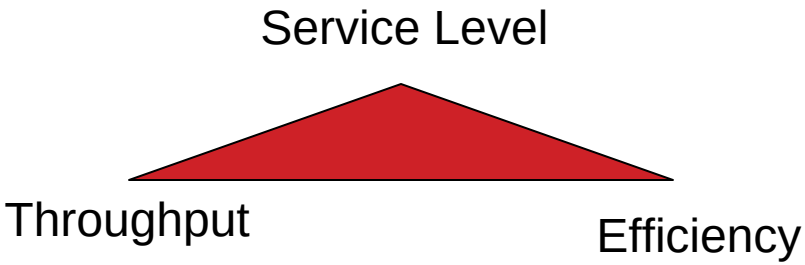
- The positioning above is about the application and data design v machine design.

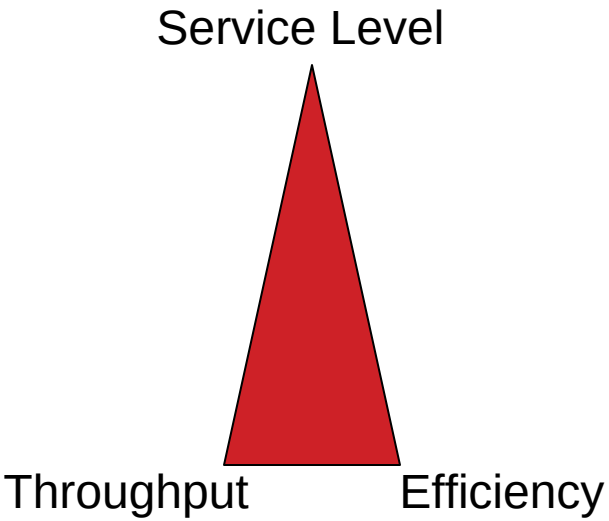
- What about the local factors?
  - Usage Pattern
  - SLA
  - Integration
  - Scale

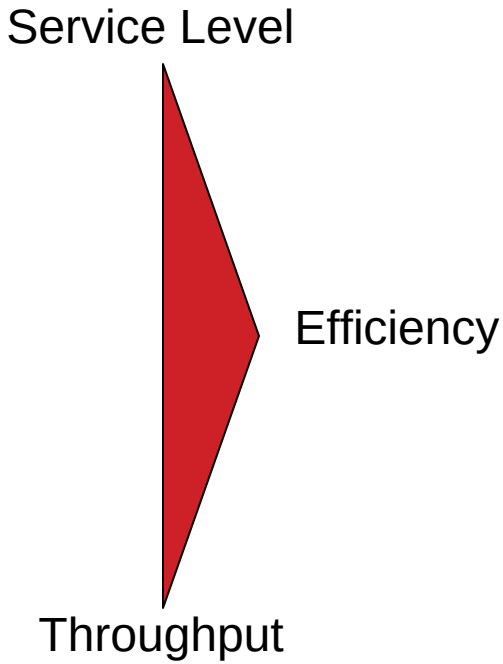
The Operational Tradeoff Triangle

The Local Factors are related to this triangle through “Normalized Headroom”







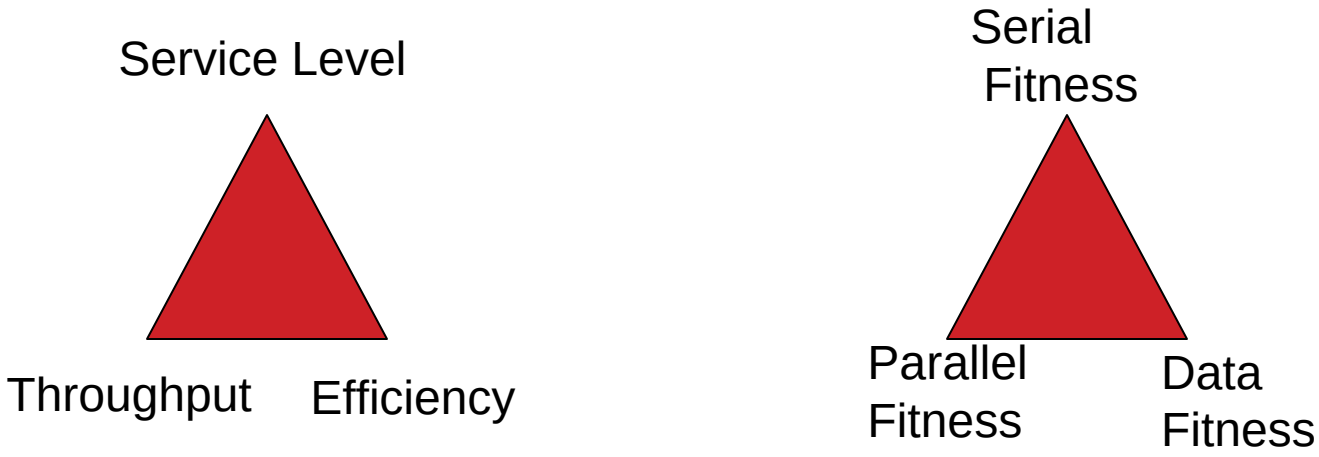


The Operational Tradeoff

Governed by “Normalized Headroom”

- $HR = (1-u)/u = c^2Nt_0/t_{wait}$
- $HR(avg) = kcN^2$
- $U = 1/(1+HR)$
- $t_{wait}/t_0 = c^2N/HR = c^2Nu(1-u)$
- $t_{wait} = (t_0)(c^2N)(u/(1-u))$
  = (capacity)(variability)(utilization)
  M/G/1 system

So how do we relate the two triangles?

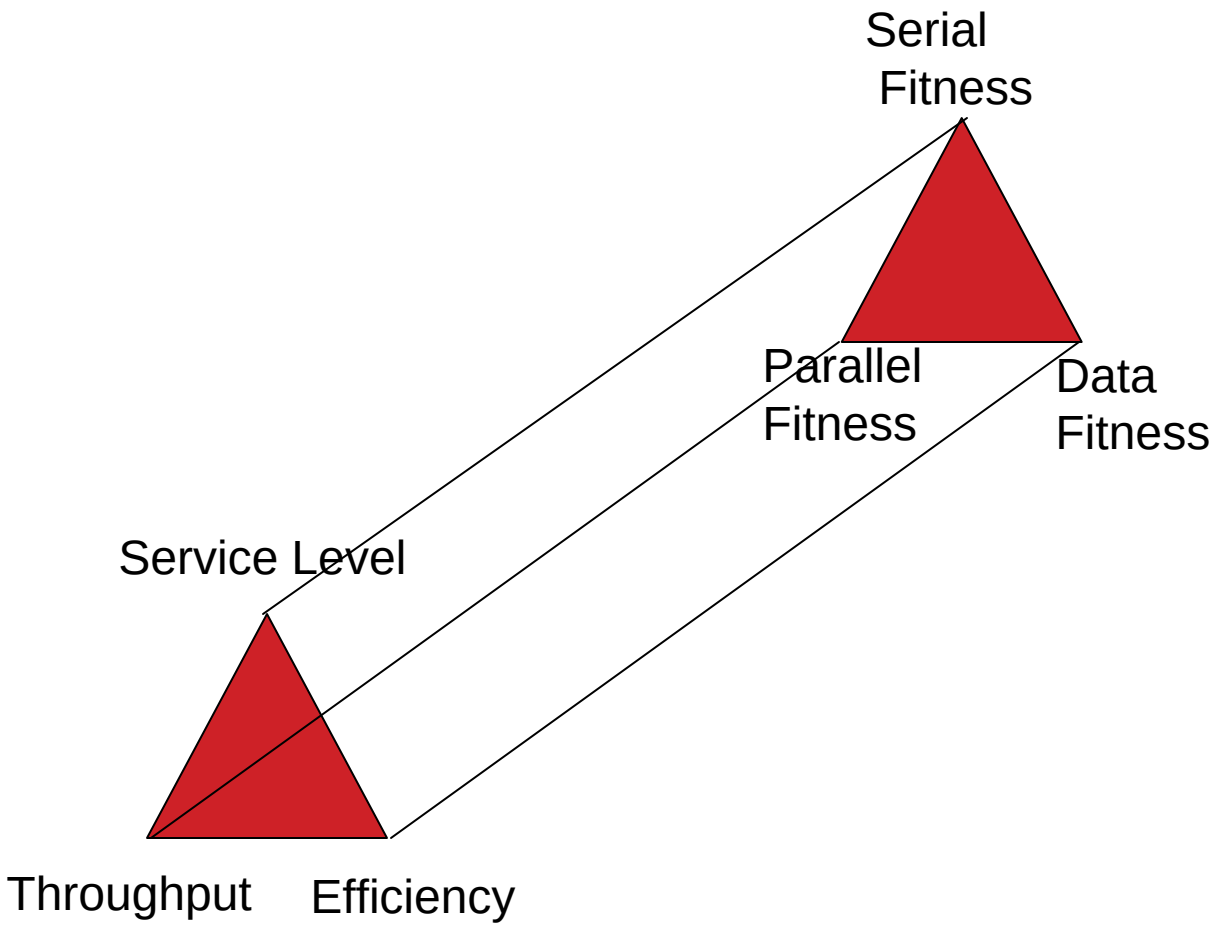


Parallel Fitness primarily drives throughput , but can drive Service Level (Wait Time)

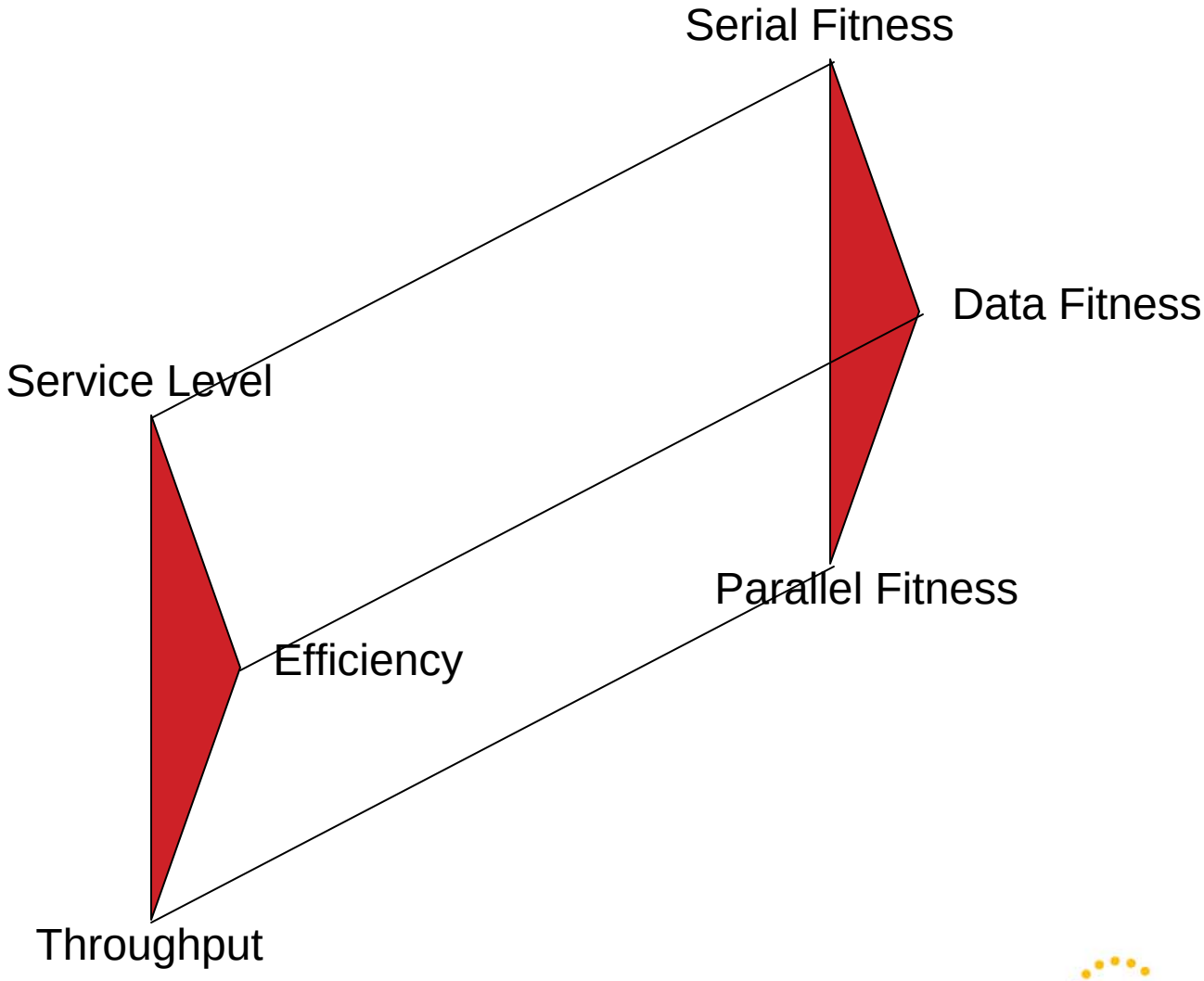
Serial Fitness primarily drives Service Level (Service Time) but will also drive throughput

Data Fitness primarily drives Efficiency by forestalling saturation, but can drive Service Level by maintaining low service time at higher loads.

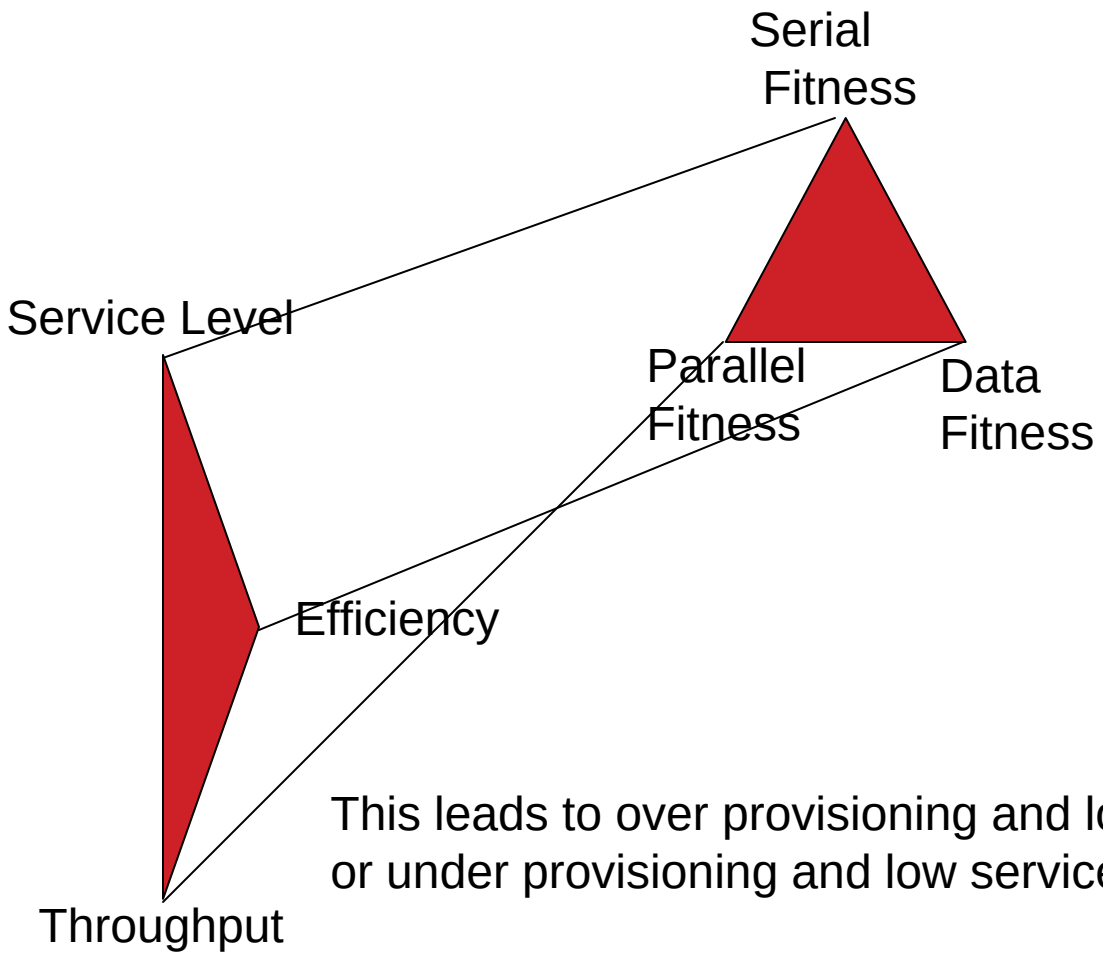
So there is a primary corner to corner relationship



The local factors distort the Client's Tradeoff

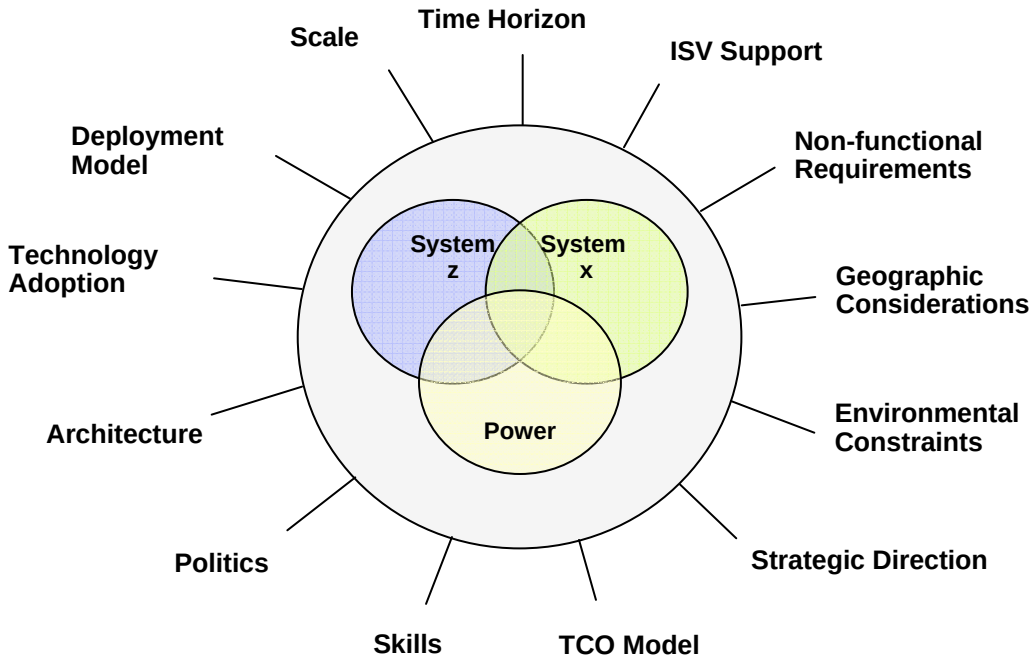


Sometimes the local factors are at odds with the design.



This leads to over provisioning and low utilization or under provisioning and low service level.

And that is just the workload optimization story



Client support is delivered through F4P Discussions, Briefings, Workshop Offerings, and ongoing System Architect support.

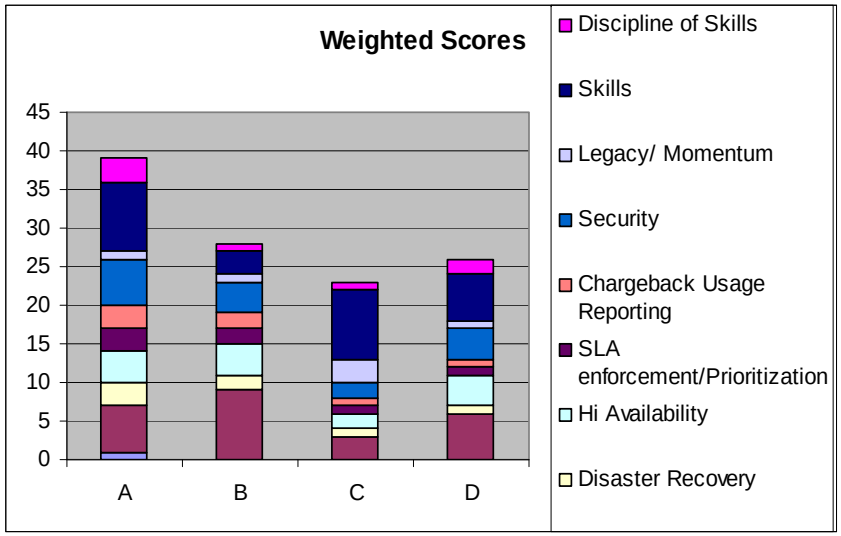
Always check to see if an SA is engaged or can be engaged for this work

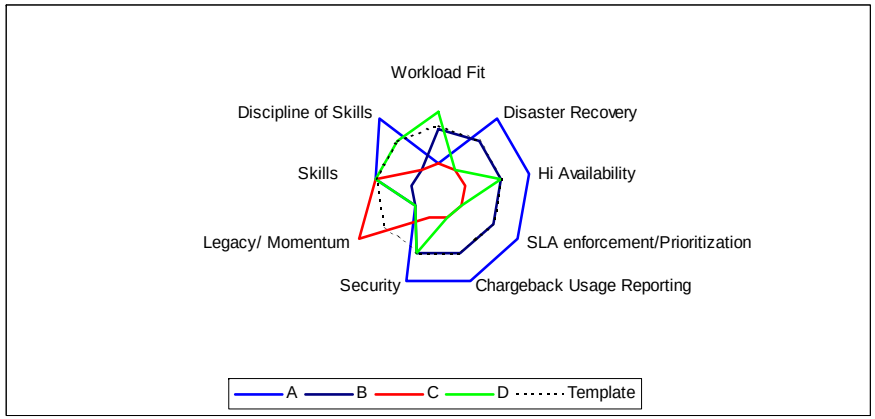
Workload is usually just one row of the F4P workshop output

Weighted Scores

Weight		A	B	C	D
3	***Workload***	6	9	3	6
1	Disaster Recovery	3	2	1	1
2	Hi Availability	4	4	2	4
1	SLA enforcement/Prioritization	3	2	1	1
1	Chargeback Usage Reporting	3	2	1	1
2	Security	6	4	2	4
1	Legacy/ Momentum	1	1	3	1
3	Skills	9	3	9	6
1	Discipline of Skills	3	1	1	2

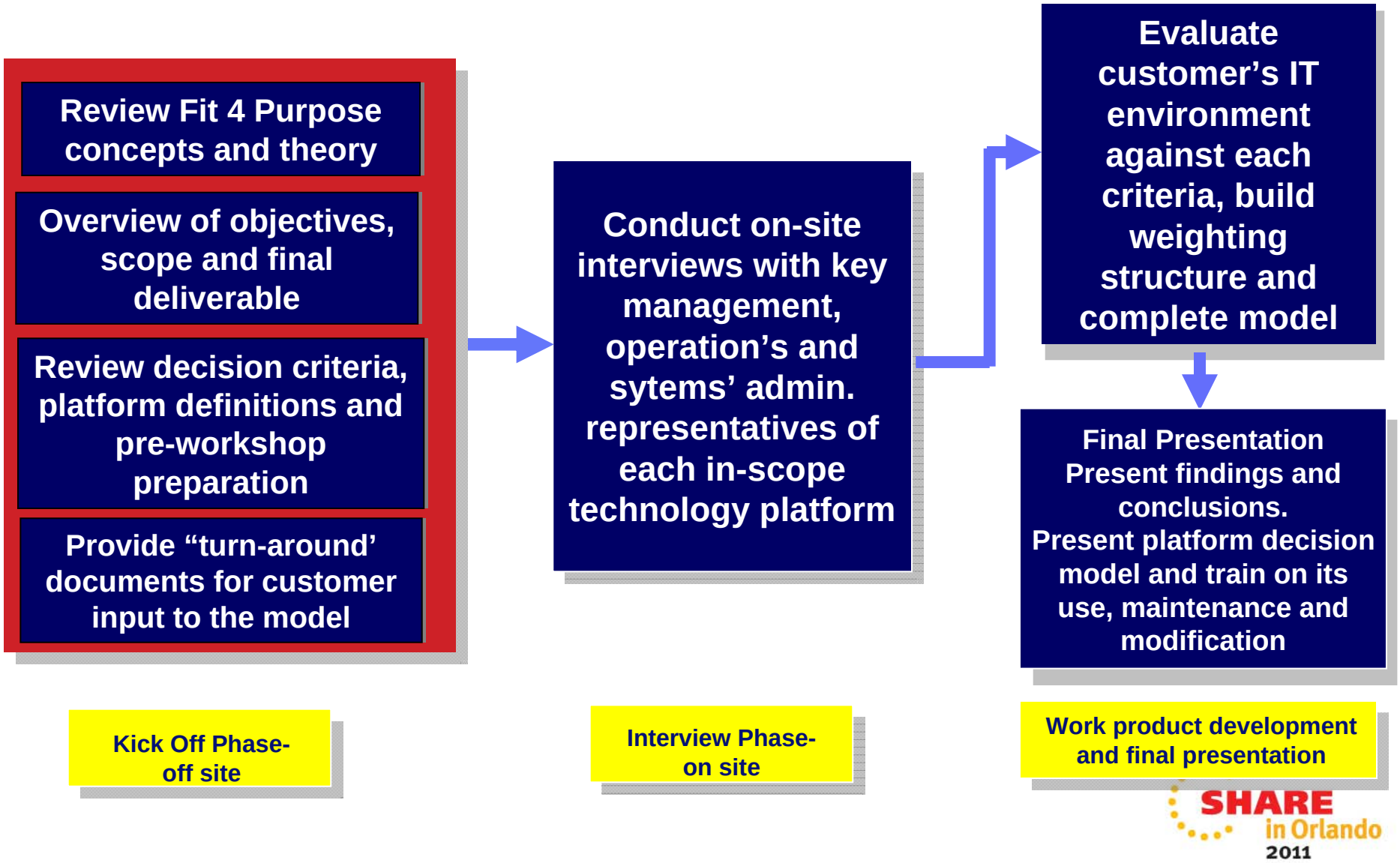
Which can yield value/requirement matching input and “foil” for TCO analysis





A provides best total value
B has best value match
C is most inline with local strategy

Lab Services Fit for Purpose workshop work flow



The lab services model: More comprehensive, precise and broader scope

Key
- Customer data entry
- Customizable
- Calculated cells; Avoid changing
- Calculated cells; Avoid changing
- Calculated cells; Avoid changing

REMEMBER TO USE A NEW COPY OF THE MATRIX FOR EACH ANALYSIS SAVING A COPY AS DOCUMENTATION.
REMEMBER TO INPUT ANY "APPLICATIONS' SPECIFIC" WEIGHTINGS ON TAB 3.

			Platform Definitions													
			Windows x86		Linux x86		Windows/Vmware x86		Linux/Vmware x86		AIX/PowerVM Risc		z/OS JAVA Risc		z/OS Cobol System z	
		Include Platform? Please Input Y or N=>		Y		Y		Y		Y		Y		Y		Y
			Platform Weights													
	Seq.	Application #1 Name	Applications' Requirements	Weight	Score	Weight	Score	Weight	Score	Weight	Score	Weight	Score	Weight	Score	Weight	Score
Decision Criteria		***Local Factors***															
	7		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	8	Application Maintainability (SW maturity, openness, etc)	5	5	25	5	25	5	25	5	25	5	25	5	25	5	25
	9	Sys. Mgt.- staff, experience, skills and tools	5	2	10	2	10	2	10	2	10	3	15	3.5	17.5	5	25
	10	Development Maturity - staff, experience, skills and tools	5	5	25	5	25	5	25	5	25	5	25	4	20	5	25
	11	Maximum Capacity	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	12		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		ISV Support															
	13	Application and platform specific vendor maturity and supportability	5	5	25	3	15	5	25	5	25	2	10	1	5	1	5
	14	Compatability with middleware, dB or other SW layer packages	5	5	25	5	25	5	25	5	25	5	25	5	25	5	25
	15		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		Workloads & System Architecture															
	16		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	17	Performance / Response Time	5	1	5	1	5	3	15	3	15	4	20	4.5	22.5	4.5	22.5
	18		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	19		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		Deployment Model															
	20	Location of Data	5	5	25	5	25	5	25	5	25	5	25	5	25	5	25
	21	Data Disaster Synchronization	5	5	25	5	25	5	25	5	25	5	25	5	25	5	25
	22	Data Synchronization	5	5	25	5	25	5	25	5	25	5	25	5	25	5	25
	25	Background or unattended	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	26		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		Scale															
	27	Enterprise Capacity	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	28	Growth (future capacity, users, etc)	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	29	Scalability	5	2	10	2	10	4	20	4	20	4.5	22.5	4.5	22.5	4.5	22.5
	30		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		Requirements & Constraints															
	31		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	32	Availability	5	5	25	5	25	5	25	5	25	5	25	5	25	5	25
	33	Disaster RTO	5	2	10	2	10	4	20	4	20	4	20	5	25	5	25
	34	Disaster RPO	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	35		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	36	Maintainability	5	4	20	4	20	4	20	4	20	4	20	5	25	5	25
	37	Data Security	5	2	10	3	15	2	10	3	15	4	20	5	25	5	25
	38	Availability - Planned outage	5	5	25	5	25	5	25	5	25	5	25	5	25	5	25
	39	System manageability and monitoring	5	2	10	2	10	2	10	2	10	3	15	3.5	17.5	5	25
	40	Data center considerations/constraints - Power	5	1	5	1	5	3	15	3	15	5	25	5	25	5	25
	41	Data center considerations/constraints - Cooling	5	1	5	1	5	3	15	3	15	5	25	5	25	5	25
	42	Data center considerations/constraints - Floor Space	5	1	5	1	5	3	15	3	15	5	25	5	25	5	25
	43		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		TOTAL SCORE			3.4		3.4		4.0		4.1		4.4		4.4		4.7

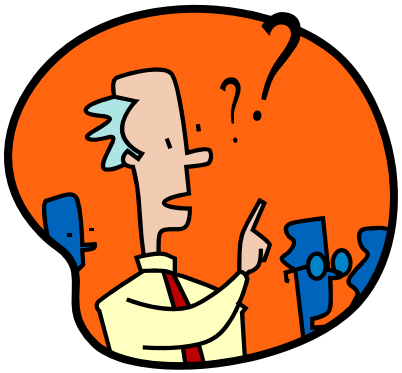
Lessons Learned

- The scope & impact will be strategic in nature.
- Relationships are key and everyone has to “behave”
  - Failed workshops have usually been caused by excessive product advocacy
- Customers like what they are seeing and are telling us they are working to determine how they will incorporate it into their own decision making process going forward.
- F4P proved to be a significant competitive advantage.
- For clients seeking answer to the question, “where do we run this application?”
  - Fit for Purpose is framework to help you decide.
  - An in depth dialog with the customer – pros and cons, tradeoffs, etc.
    - *Not a sales call… consultative*
    - *Identify value differentiators and weightings*
    - *Ongoing reuse of “local factors” climbs a learning curve*
- Indeed one size does not fit all; we have both mainframes and “Watson” for good reasons.

Key Challenges

- Being Client Centric in a brand dominated environment
  - Requires a strong minded client first approach
  - Avoid advocacy: F4P is more about “listening” than “telling”
    - There is a fine line between articulating value and platform advocacy
    - The client defines the line by the questions asked and assertions made

- Avoiding our own brand centric biases
  - We all come with baggage
  - Many things that “everybody knows” are not true
    - Most of what you know about relative capacity falls into this
  - Avoid accepting the common wisdom about “the once and future platform”
  - Need to avoid gaming the system unintentionally
  - Can’t tolerate those who game the system intentionally
  - AHP method can help
    - Pairwise comparison of requirements and platforms
    - Can’t see overall picture until the end
    - Hard to apply spin until done.



Questions

Some “Light Reading” on the Topic

- “Server Platform Selection and Positioning”, Lebsack, Dixon CMG Conference Proceedings, December, 2009..
- Dr. Gregory Pfister , In Search of Clusters, the ongoing battle in lowly parallel computing, Second Edition, Prentice Hall, 1998
- Roger Rogers and Joe Temple, “Relative Capacity and Fit for Purpose Platform Selection”, CMG Journal of Computer Management, no 123, March 2009
- Dr. Neil Gunther: Guerilla Capacity Planning
- Rick Lebsack and Joe Temple “Fit for Purpose Platform Selection, a Workload View” on IBM Techdocs and pending CMG Journal
- See Joe Temple for drafts: “The Operational Trade off Triangle”, “The Server Design Trade off Triangle” and “Using Normalized Head Room for Infrastructure Analysis and Design”