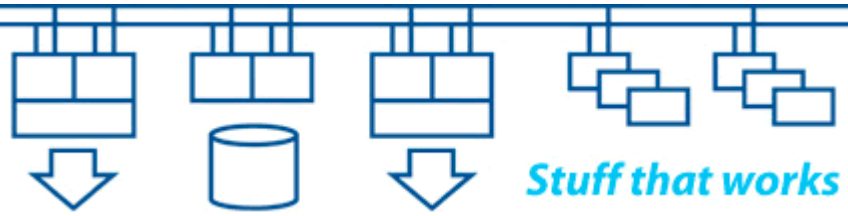
UKCMG Free Forum

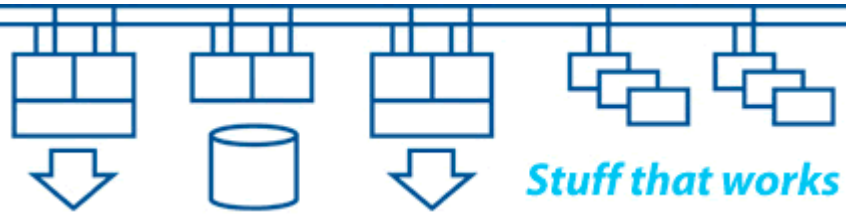
How performance affects service

Colin Butcher



What is "service"?

Service provision is about meeting the realistic needs and expectations of the users for access to the systems and data within the constraints that are placed around the design, implementation and operation of a system by the business or organisation.



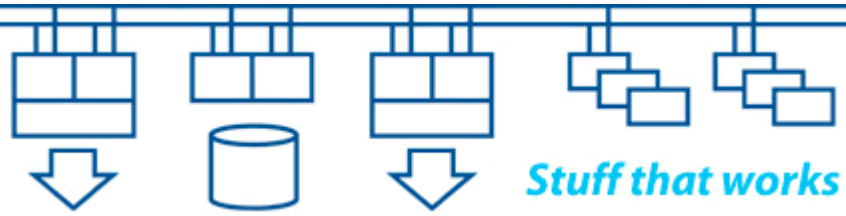
What is "performance"?

Performance management and capacity planning are about ensuring that a system behaves predictably, is stable and remains available for use while coping with whatever is reasonably thrown at it in terms of workload and data volume.

We also have to ensure that the surrounding network and storage infrastructure is equally capable in terms of performance, capacity and availability.

Overview

- Failure to perform is equivalent to loss of service
- Performance related failures tend to be transient and difficult to reproduce
- Failures may show up some time after the problem occurred, for example data corruption
- How do you want the system to fail?
- How can we reproduce a problem in order to understand it, fix it and prove that it's been fixed?



The performance envelope

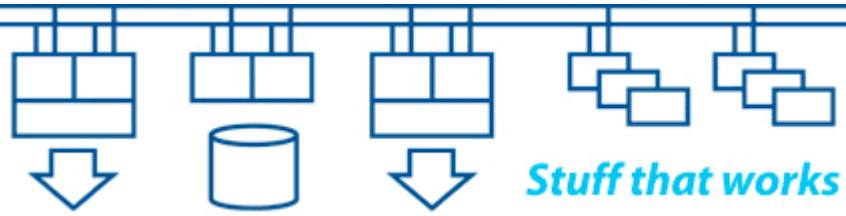
- What do the systems have to do?
- What happens if the systems fail?
- What happens if you push beyond the limits?
- How far from the edge are you?
- How do you know?

How can we characterise a system?

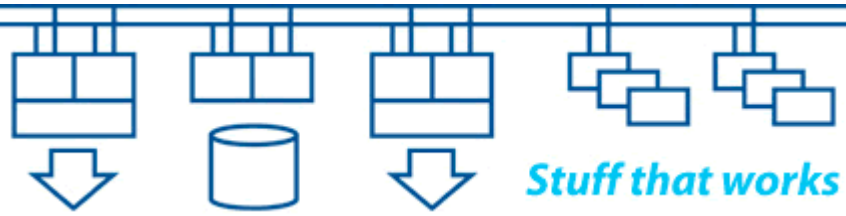
- What is "fast" or "slow"?
- What can we measure?
- What comparisons can we make?
- What "footprint" can we look for if there is a problem?
- How can we track down where a problem is?

- **Bandwidth – determines throughput**
  - It's not just "speed", it's throughput in terms of "units of stuff per second"
- **Latency - determines response times / round trip delays**
  - Affects how much "stuff" is in transit through the system at any given instant
  - "Stuff in transit" is the data at risk if there is a failure
- **Jitter ("div latency") – determines predictability of response / round trip delay**
  - Ideally close to zero, ie: predictable response / round trip delay times
  - Severe latency fluctuations can cause system failures

- Understand scalability – do as much as possible once only, do little as possible every time
- Understand how the applications could break down into parallel streams of execution:
  - Some will be capable of being split into many small elements with little interaction between the parallel streams of execution
  - Some will require very high interconnectivity between the parallel streams of execution
  - Some will require high-throughput single-stream processing
- Get the "state machine" design right for multiple parallel streams of execution (multi-threading) – look for timing windows and race conditions

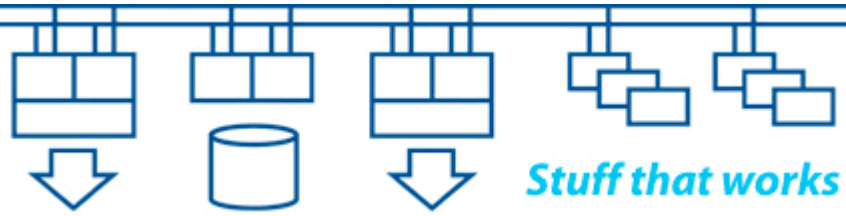


- Ability to make use of hardware parallelism
- Synchronisation and serialisation techniques
- Scalability techniques
- Designing and writing very good code requires very good programmers



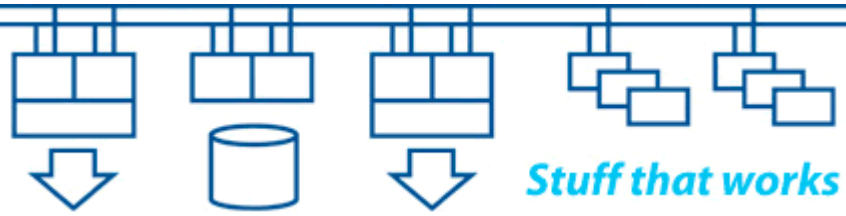
What can we do?

- Understand the workload and size systems accordingly
- Scalability – do as much as possible once only, do little as possible every time:
  - The fastest IO is the IO you don't do
  - The fastest code is the code you don't execute
- Look through the entire network infrastructure that connects the users and data feeds to the systems
- Look through the entire storage infrastructure that holds, replicates and backs up your data



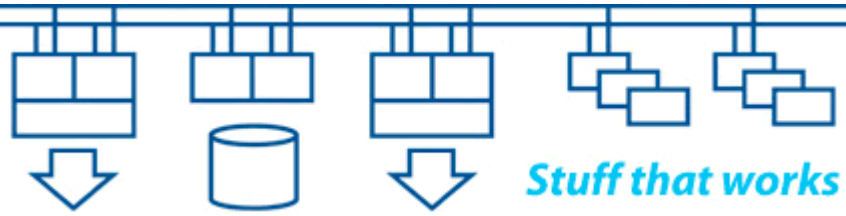
How can we test it?

- We must test for scalability and performance, not just functionality
- How hard can we push a system before it fails?
- What data can we use for testing?
- How can we simulate all the external interfaces and data flows in a realistic manner?

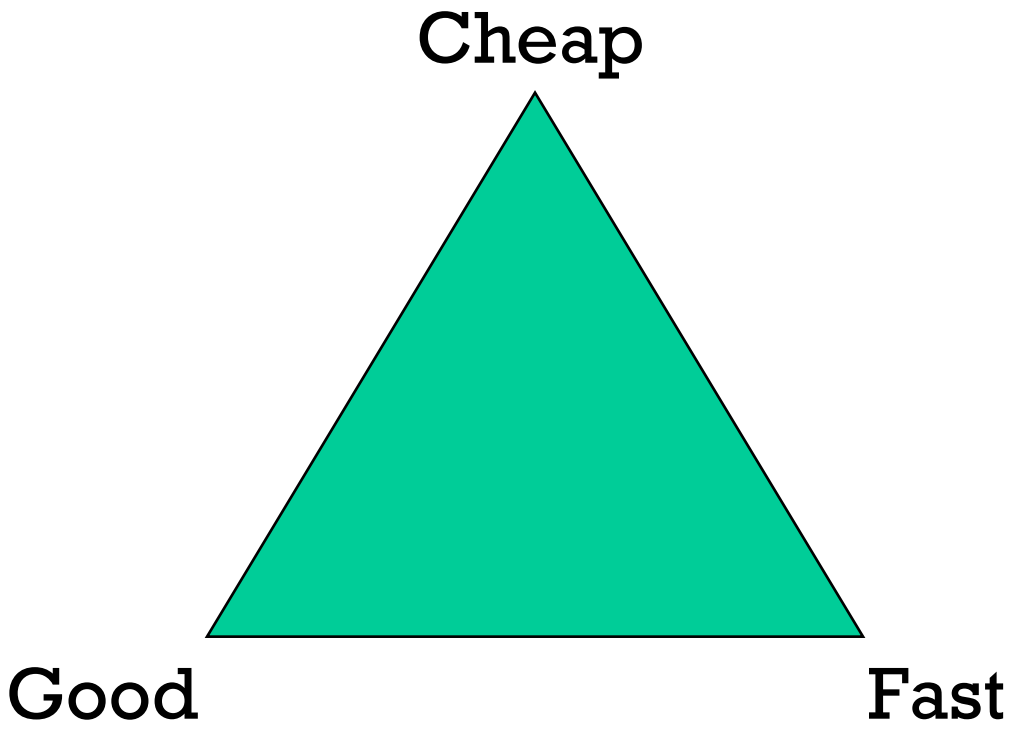


Summary

- We need systems to be capable of absorbing unexpected spikes in workload without problems
- We don't want to spend our time managing performance and doing tuning exercises
- Designing the entire systems and surrounding infrastructure for good performance and scalability on multi-processor systems is key to successful operation
- We have to do realistic testing before we put it into live production



Compromises – pick any two!



Finally – don't forget the people

In order to deliver good service we have to ensure that the support and operations staff receive appropriate commitment from the business or organisation…

Co-operation, Training, Respect, Funding, etc.

… all of which have to be earned by delivering good service and communicating with the rest of the business or organisation on a regular basis.

Thank you for your participation

Q & A