
# SLAM DUNK. How to score full points on performance in SLAs.

**Adam Grummitt – adam.grummitt@metron.co.uk**

Director
Metron

# Abstract: SLAs and Performance Assurance

- SLAs define IT service requirements formally
- Constrain/contract both receivers & providers
- Define/repository for Performance Targets
- Measurable key performance indicators (KPI)
- Business and application views inc. response time
- Models used to reflect and police SLAs
- Establish a performance management regime
- Threshold violations alarms and alerts
- Achieve Performance Assurance

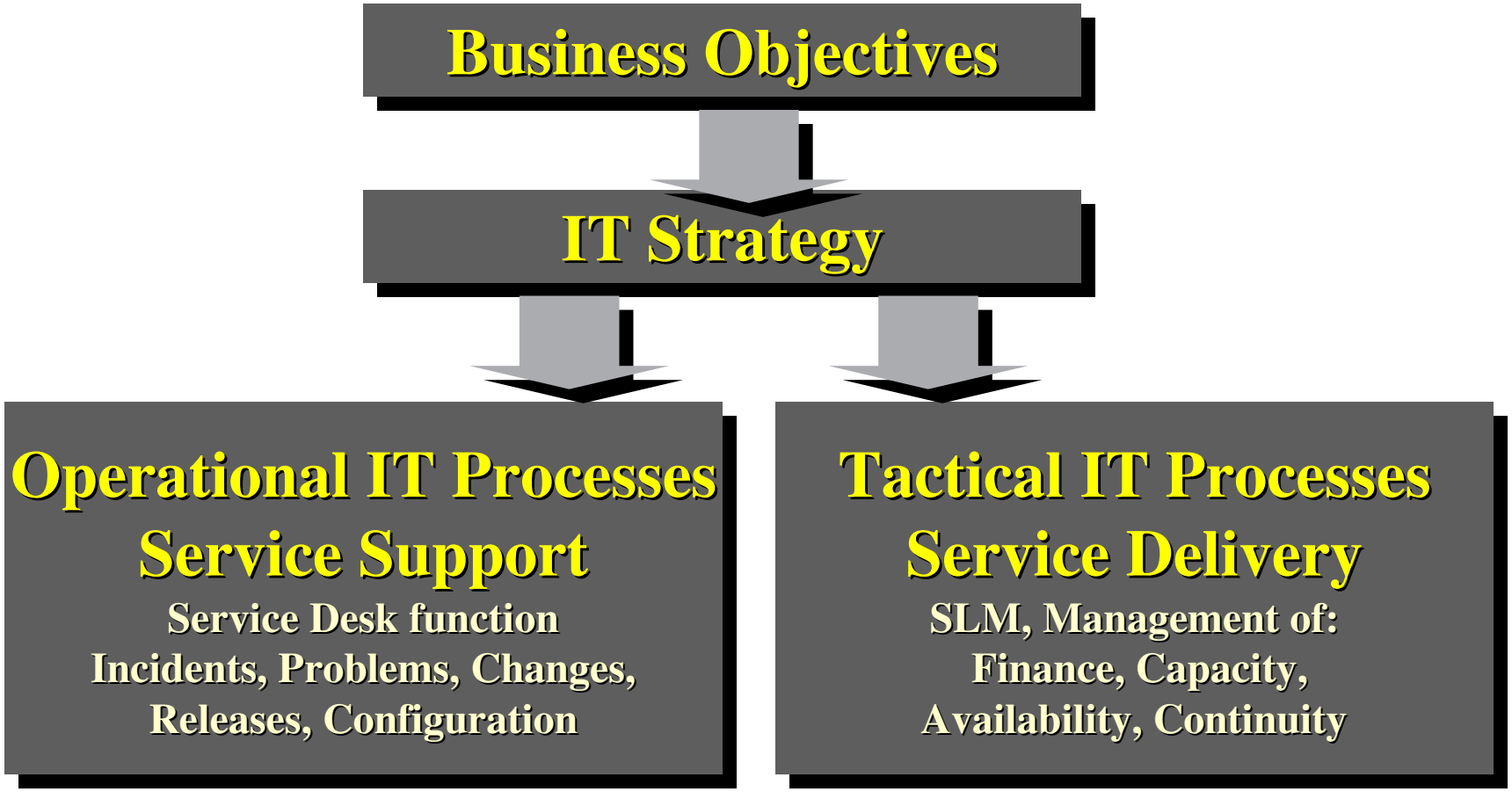
# Introduction

- SLA & SLAM – ITIL & ITSM
- Six Sigma, BSC, CoBit, QM, MOF
- Practical approach to performance in SLAs
- A skeleton SLA
- Typical outcome analysis
- Use of capacity management techniques
- Typical implementations and benefits
- Samples including ecommerce

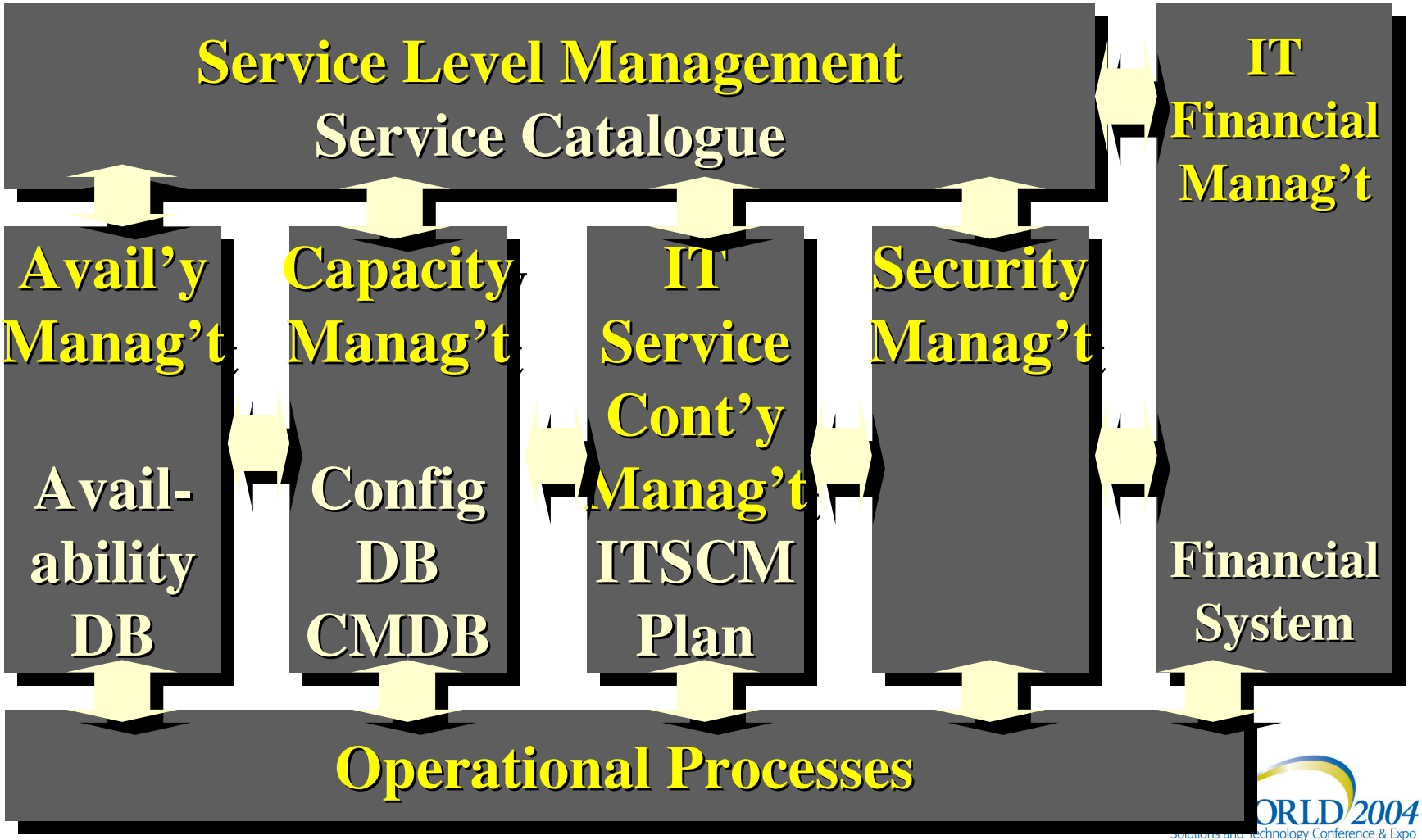
# ITIL

- The ITI Library - books & definitions
  - Service Support & Service delivery
  - Business, Infrastructure, Development, Service
- Good practice for managing IT
- Basis of BS15000, 7799 and ISO 17799 standards
- Developed by UK's OGC in the 90's
- Metron key contributor to initial Demonstrator
- itSMF
  - The IT Service Management Forum for ITIL users
  - Promotes exchange of info & experience
  - GB, NL, B, AUS, ZA, CDN, F, CH/A/D, USA

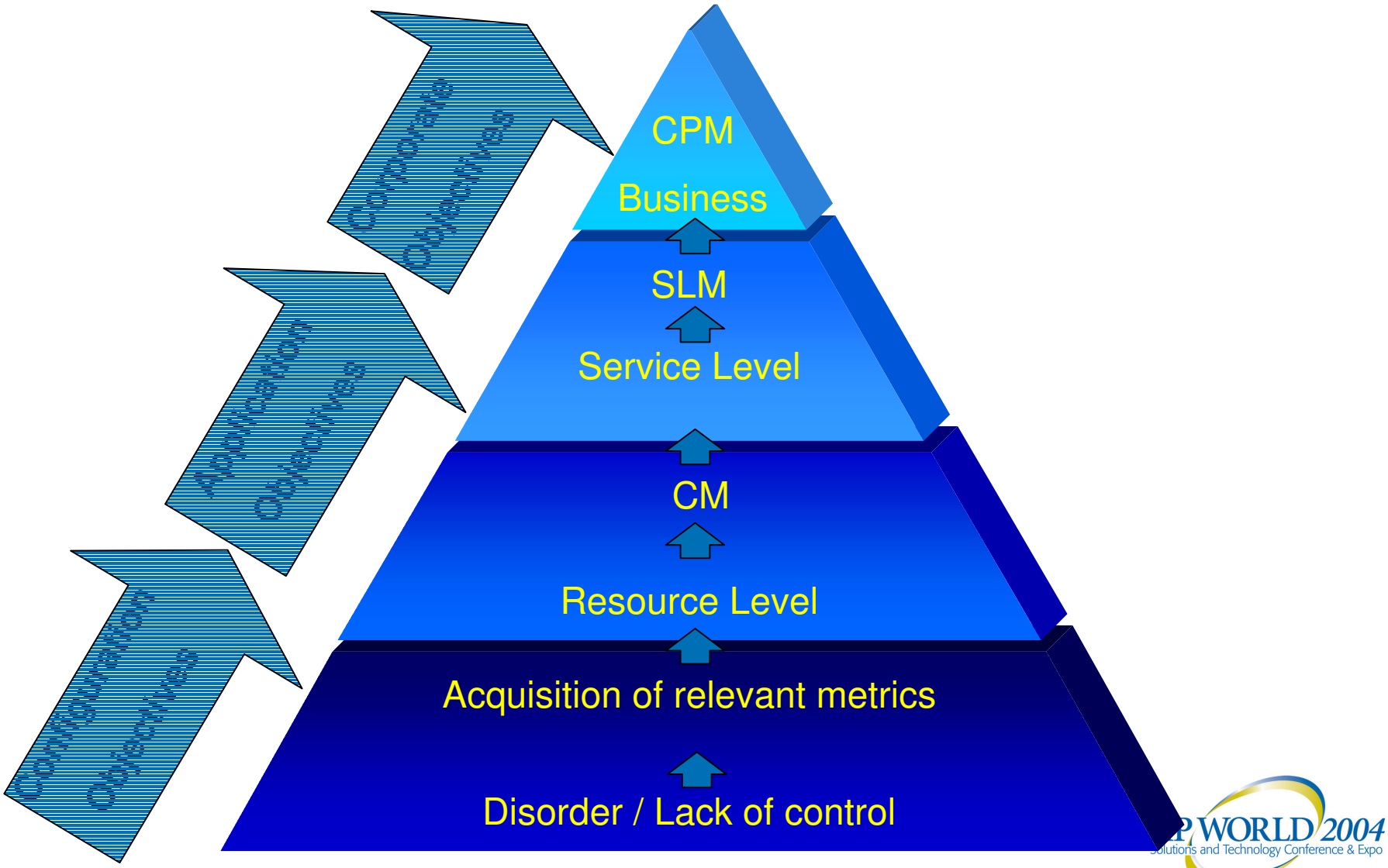
# ITIL overview



# ITIL Service Delivery Processes



# Performance Process Pyramid



# Options

Performance Assurance

Assess improvements (EFQM)

Operate (6σ - DMAIC)

Implement CM (ITIL)

Plan attack (FMEA)

Understand target (SIPOC)

Performance metrics (COBIT)

Route rationale (FMEA/Lean)

Current position (ITIL/BS15000)

Decide on business objective

Aim for total quality (TQM)

People, Process (BSC)

Instal agents & create PDB

Define CPM (CMM)

Define business metrics

Define data capture strategy

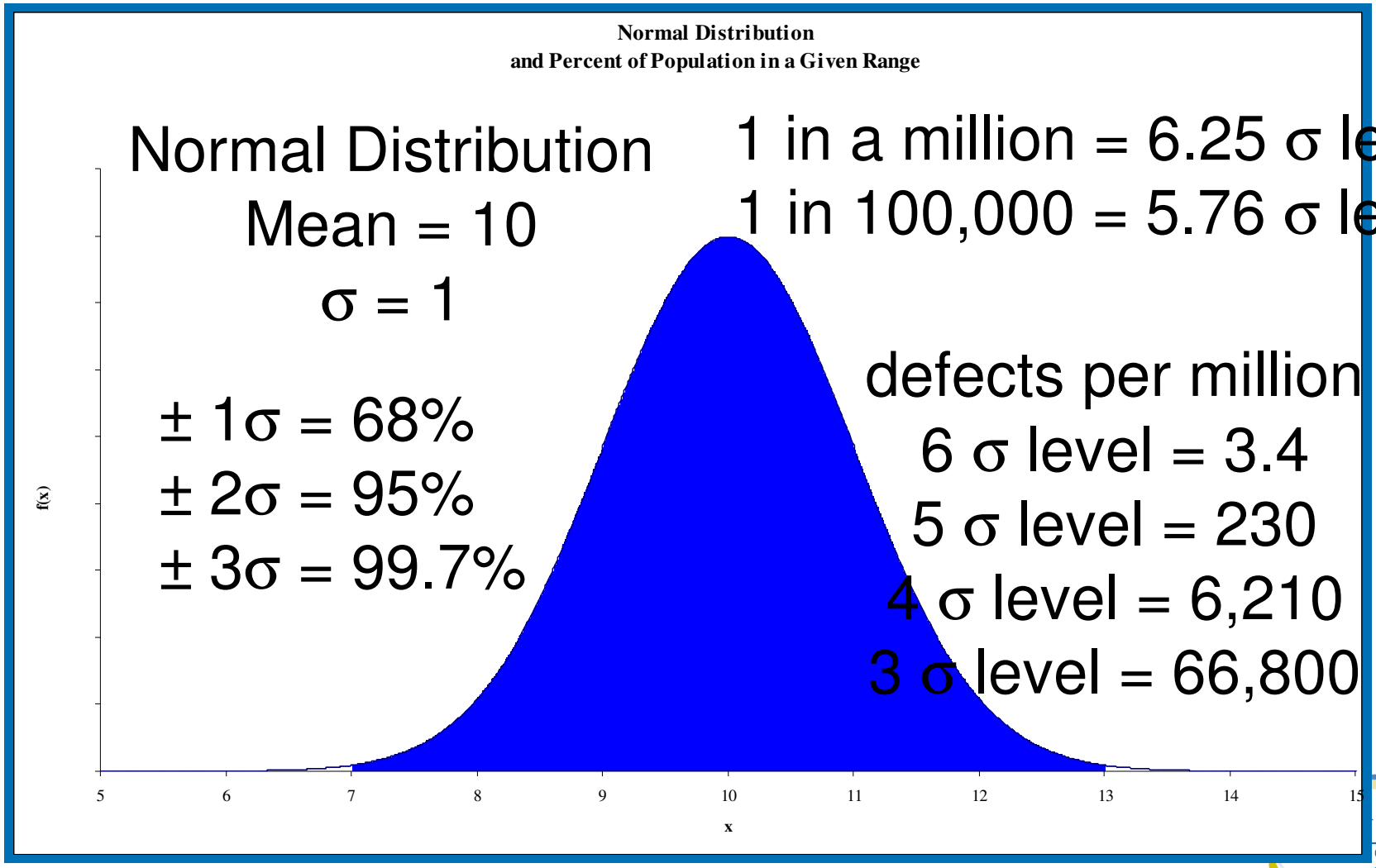
Assess position (BS15000)

Understand customers (VoC)

Business req's (CtQ)

Performance chaos

# Six Sigma


Normal Distribution
and Percent of Population in a Given Range
Normal Distribution
Mean = 10
σ = 1
± 1σ = 68%
± 2σ = 95%
± 3σ = 99.7%
1 in a million = 6.25 σ level
1 in 100,000 = 5.76 σ level
defects per million
6 σ level = 3.4
5 σ level = 230
4 σ level = 6,210
3 σ level = 66,800
f(x)
5
6
7
8
9
10
11
12
13
14
15
x

# BSC EG

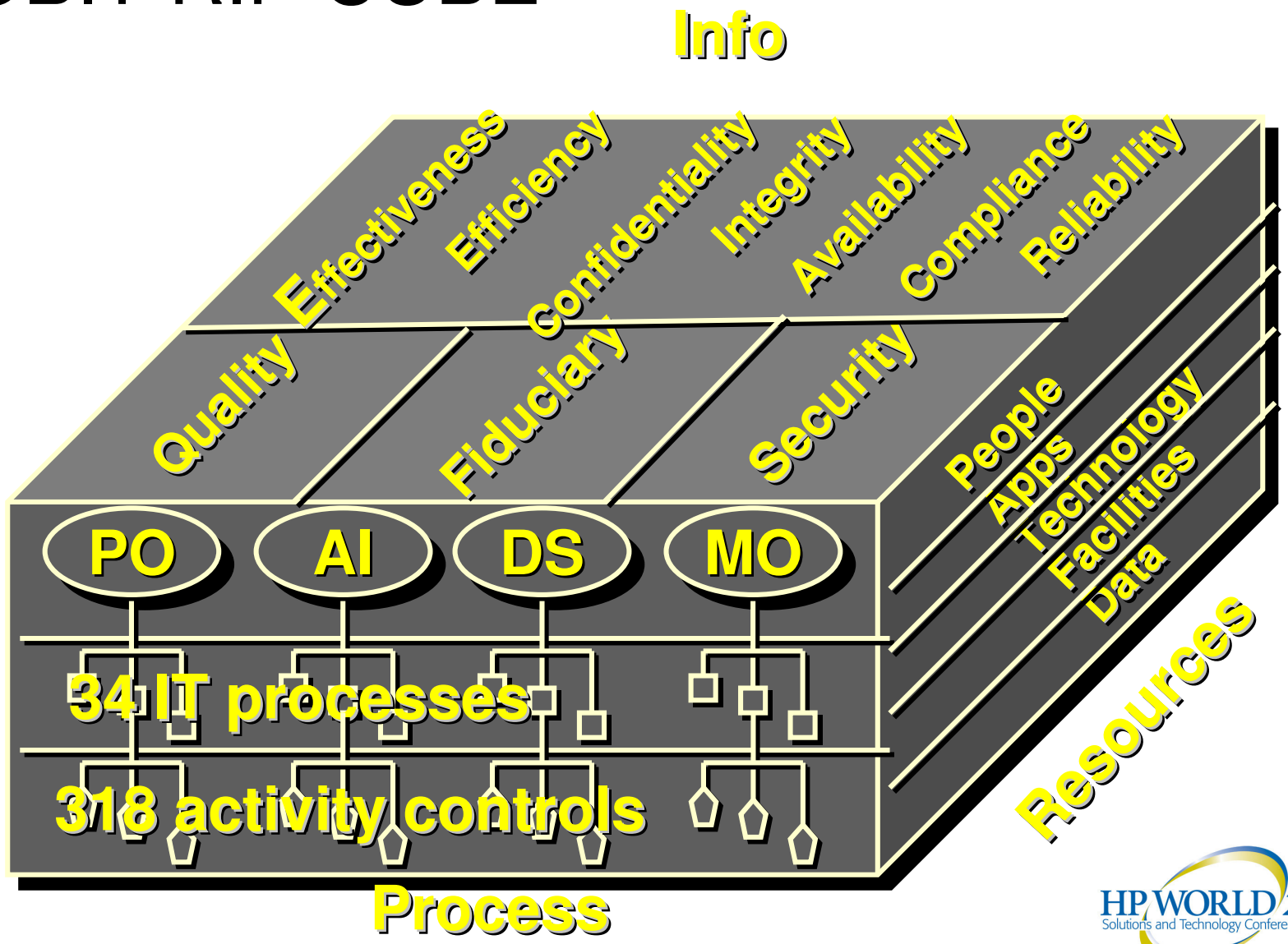
| Process | KPI | # | | σ |
|---|---|---|---|---|
| Finance | Cost/income | 7.3 | | 2.1 |
| | Profit v Plan | 4.2 | | 1.8 |
| | Risk exposure | 3.2 | | 1.6 |
| | Earnings/share | 8.8 | | 1.2 |

| Process | KPI | # | | σ |
|---|---|---|---|---|
| Internal | Staff turnover | 4.3 | | 1.1 |
| | Unit trans cost | 9.1 | | 0.7 |
| | Risk exposure | 4.1 | | 1.0 |
| | Conformance | 9.4 | | 0.3 |

| Process | KPI | # | | σ |
|---|---|---|---|---|
| Client | Intermediation | 6.1 | | 1.2 |
| | Win rate | 3.6 | | 0.2 |
| | Market share | 4.3 | | 0.3 |
| | Satisfaction index | 8.8 | | 1.2 |

| Process | KPI | # | | σ |
|---|---|---|---|---|
| New | Product Sales | 2.3 | | 1.1 |
| | Sales/employee | 5.2 | | 2.7 |
| | Staff Attitude | 3.4 | | 1.6 |

# COBIT RIP CUBE

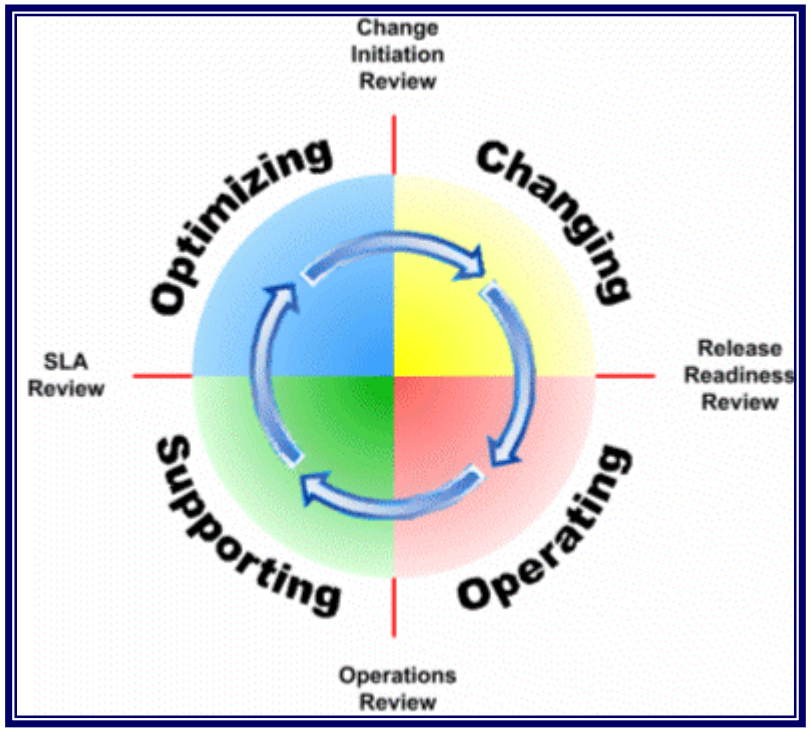


# CMM, TQM, EFQM

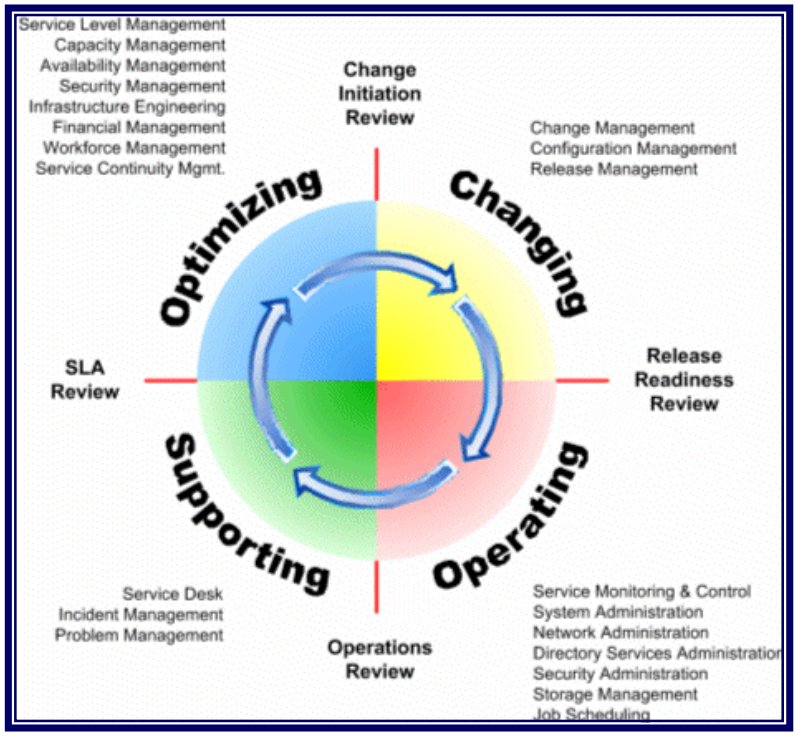
- CMM levels and QM

1. Basic introduction, ad hoc solution
2. Repeatable or reactive process
3. Defined or proactive process
4. Quantitatively managed process
5. Optimised effective implementation

# MOF




Process model




Process model
& ITSM functions

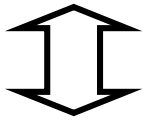
# SLAs

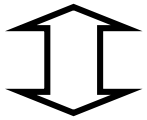
- Quantify obligations of provider & receiver
- More important if services externally charged
- Functions that the service will provide and when
- Need measurable performance indicators
- Mutual interest for it to be clear & measurable

# SLAs & Capacity Management

QA ⇔ Capacity Management ⇔ SLA

(Performance Assurance)





## Performance Management

- Resource accounting
- Workload balancing
- Program optimisation
- System tuning
- Alarms and alerts
- Reporting
- Tracking

## Capacity Planning

- Application sizing
- Workload trending
- Workload characterisation
- Performance Forecasting
- Modelling
- Reporting
- Tracking

# SLA processes

- Measurable numbers > arbitrary guesstimates
- Assess system at early stage in its production life
- Granularity of models ∝ questions to be answered
- Split total workload into workload components
- “What-if” scenarios to assess likely bottlenecks
- Results identify thresholds for monitoring metrics
- Web reporting system - automatic alerts & alarms

# SLA Skeleton

- Scope - parties, period, responsibilities…
- Description – application, what is (not) covered
- Service hours – normal, notice for extension…
- Service availability – % uptime in defined periods
- Service reliability – usually defined as MTBF
- User support levels – MTT respond/ resolve/ fix
- Performance – throughput, responses, turnaround
- Minimum functionality – basic service
- Contingency – continuity, security, standby
- Limitations – agreed restrictions on usage
- Financial – charging, incentives, penalties…

# SLA iceberg

- Hardware on which the system will run
- Traffic incurred
- Other workloads on the same machine
- If app on another machine/test, then measure it
- For new apps in particular, workload trials in QA
- Definition of a workload and what to measure
- Emulation or replication or a controlled workload
- If app is in development, then use SPE

# SLA & Performance

Typical

- “Mandatory response of 3 secs; desirable 1 sec”
- “Mandatory 8 secs; desirable 5 secs for 95th %”
- “Normal maximum peak traffic of 3,600 per hour”
- “Normal service regime for 08:00 – 18:00”

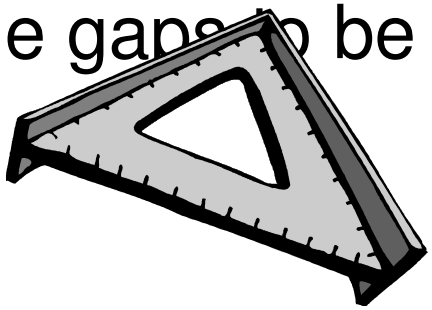
But

- Need measures that can be monitored and used
- Spurious statistical detail re uniform distributions
- “Twice the standard deviation”, 95th percentiles

These are all part of Capacity Management

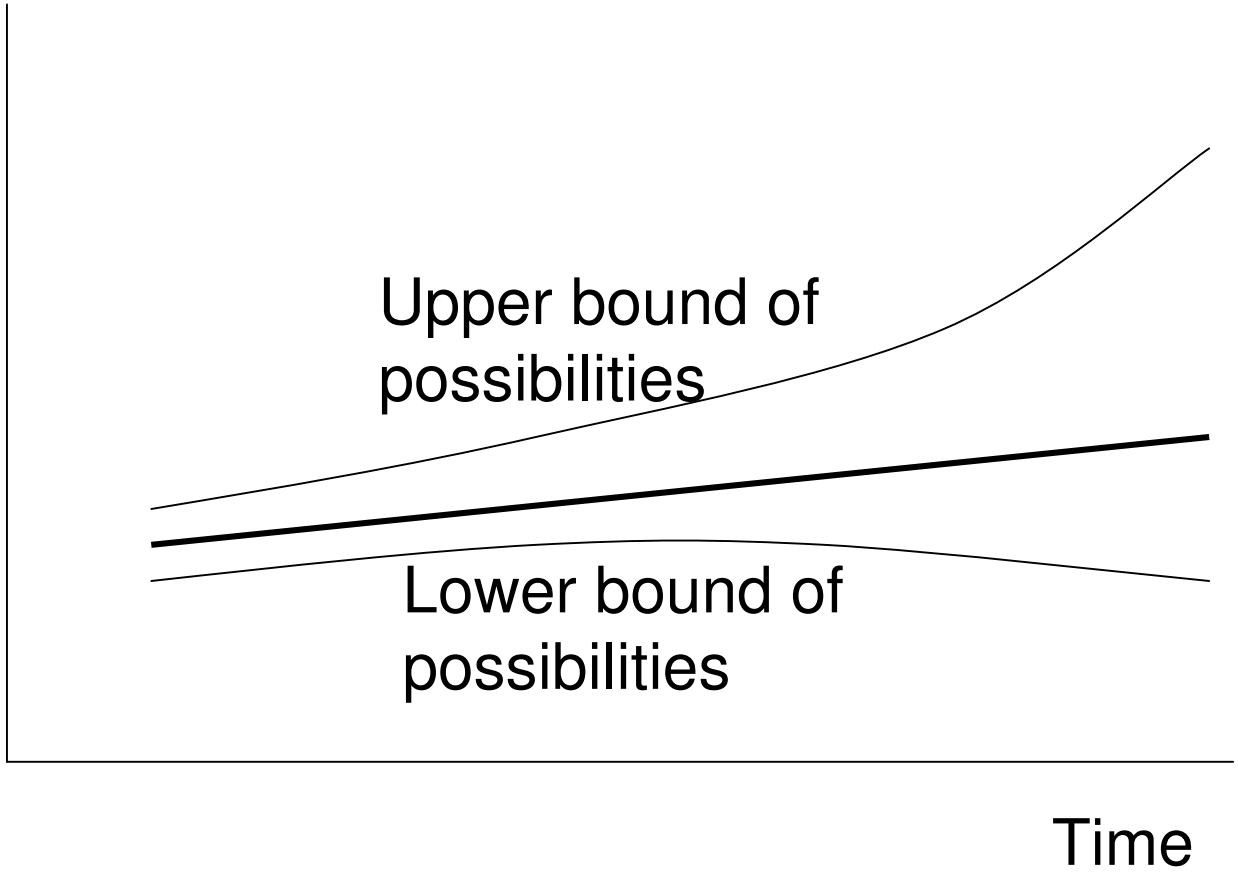
# Performance Metrics variability

- Metrics are variable in presence and reliability
- What is available is not always necessary
- What is necessary is not always available
- Both system level and user/process level
- Metrics may be sparse re mapping or responses
- Some applications are well instrumented…
- Network statistics mostly in ports, packets…
- Rules and practices enable gaps to be filled

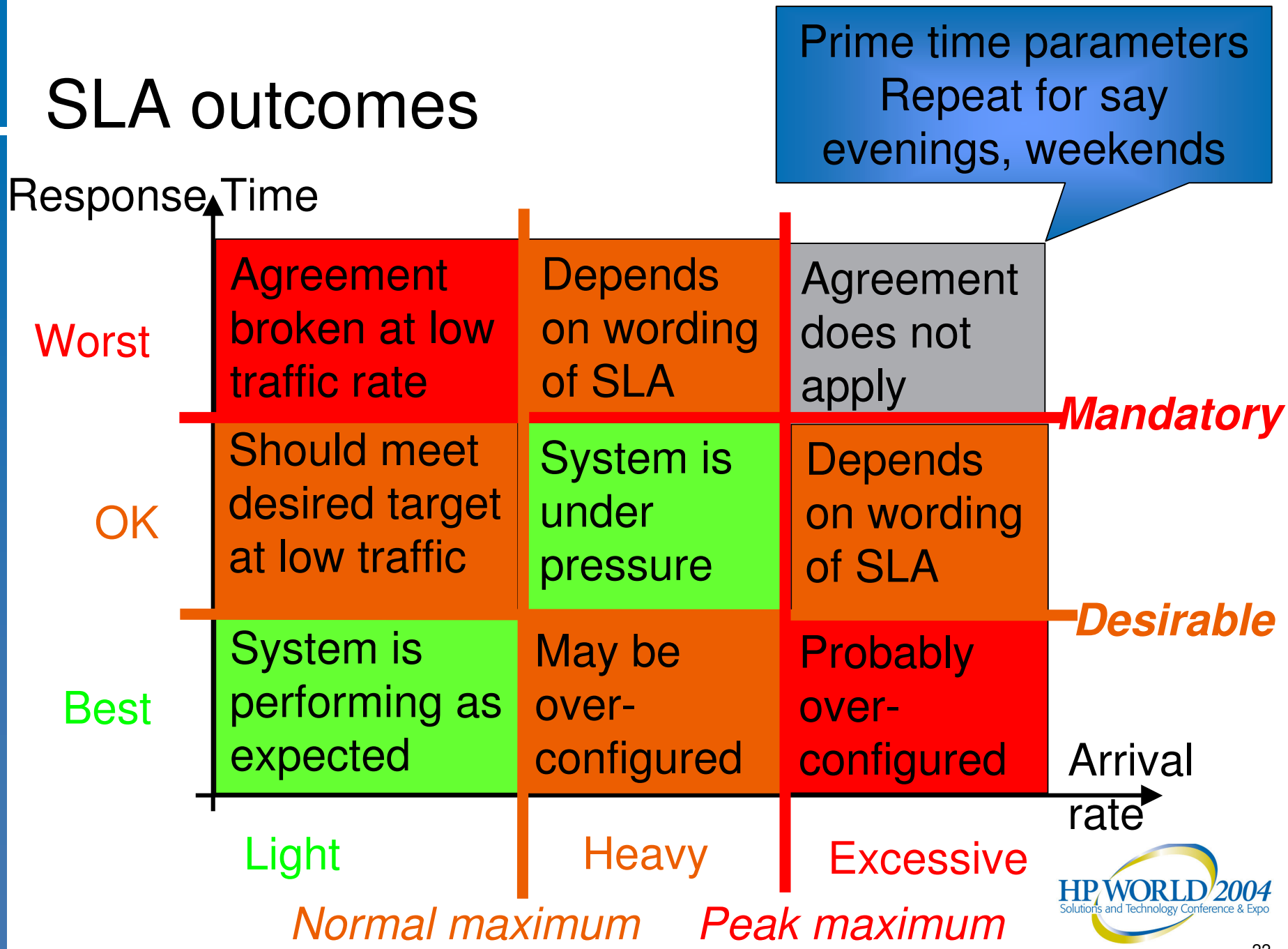


# Sensitivity Analysis

Parameter Being forecast e.g. arrival rate, utilisation



# SLA outcomes



# Capacity Management & SLAM

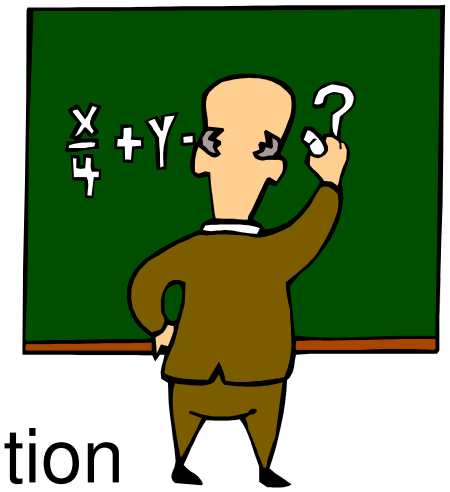
A framework for building SLA performance:

- Characterisation of workload components
- Evaluation of SLAs via modelling tools
- Reporting by workload components
- Automation of monitoring and reporting
- Automation of alerts/alarms on violations

- Monitoring the present
- Analysing the past
- Predicting the future

# Analytic Model assumptions

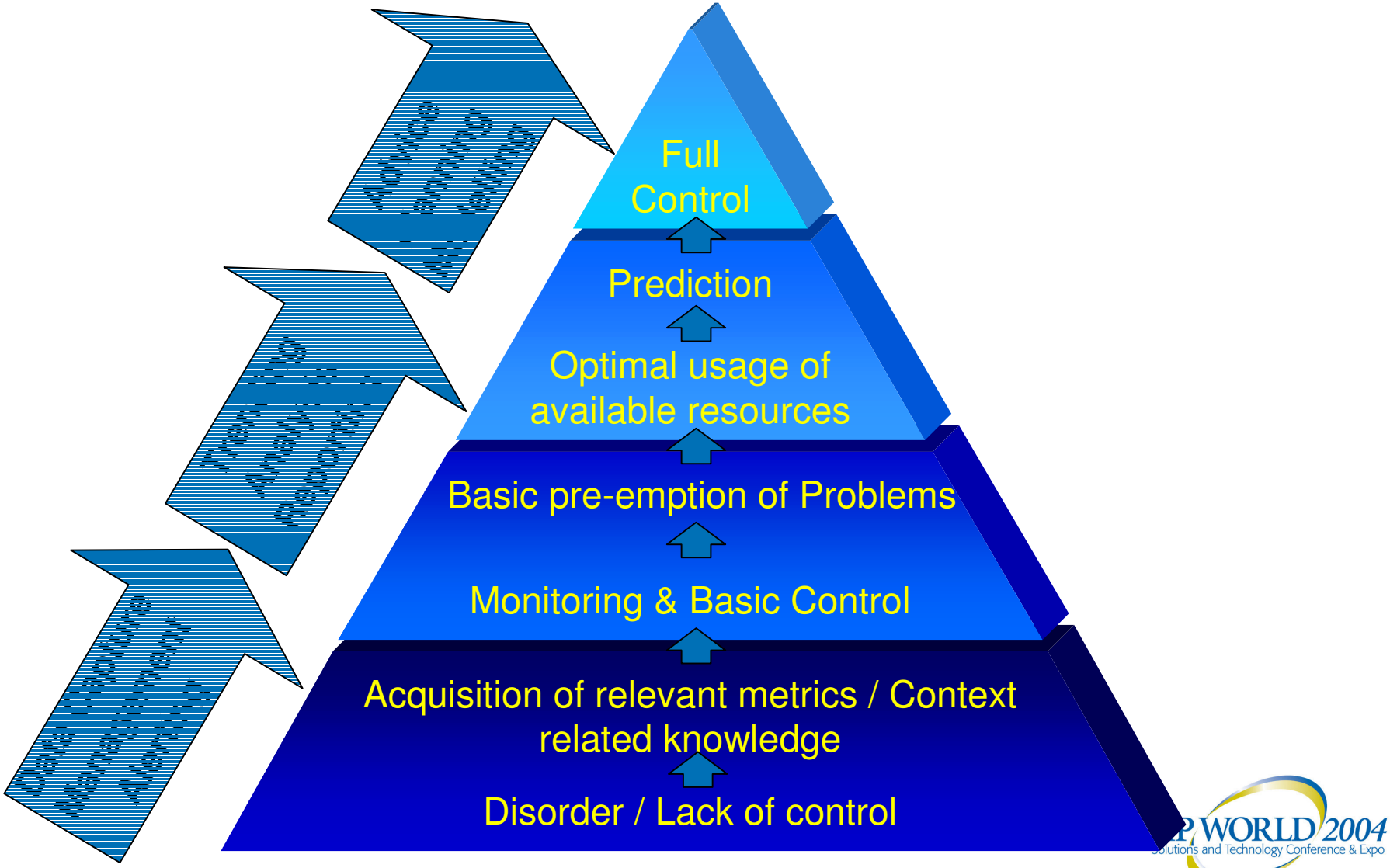
- Use multi-class queuing network theory
- Assume large populations of transactions
- Assume exponential distributions:
  - Service times
  - Inter-arrival gaps
- "Typical" transaction is an average
- Typical SLAs assume normal distribution
- The 95th percentile usually taken as $2\sigma$

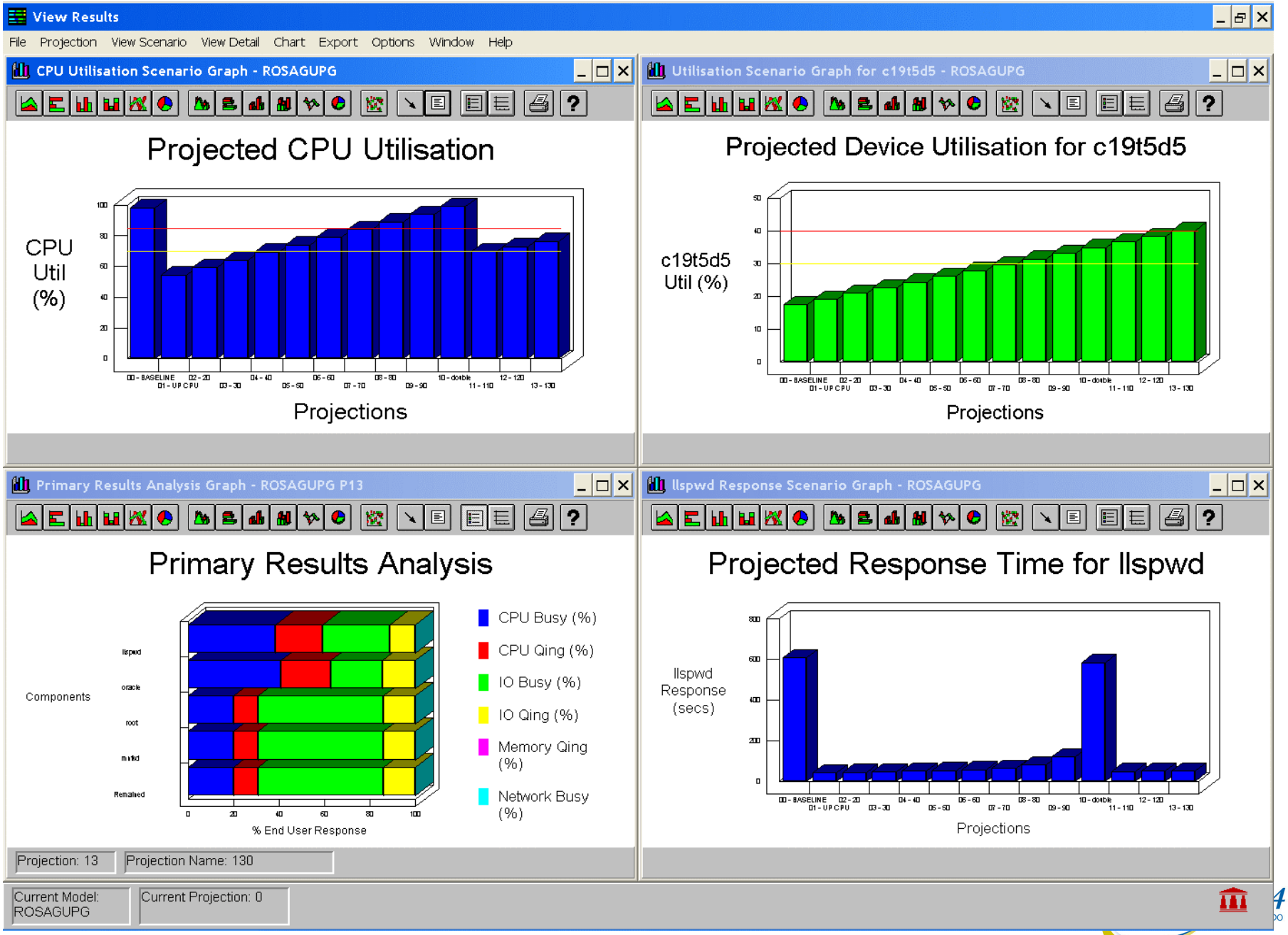


# Performance Assurance tools

- SLA definition of an app depends on the site
- Typically, n users all running a particular package
- A large number of transactions via an even larger number of processes
- Need to capture, collect and store all KPI details
- Aggregate all the resource demands for a group of processes or users = workload component
- Synthesised - usually not a “real” transaction
- Used to define a baseline situation and assess relative degradation with increasing traffic etc.

# Performance App Pyramid




View Results
File Projection View Scenario View Detail Chart Export Options Window Help
CPU Utilisation Scenario Graph - ROSAGUPG
Projected CPU Utilisation
CPU Util (%)
Projections
Utilisation Scenario Graph for c19t5d5 - ROSAGUPG
Projected Device Utilisation for c19t5d5
c19t5d5 Util (%)
Projections
Primary Results Analysis Graph - ROSAGUPG P13
Primary Results Analysis
Components
% End User Response
CPU Busy (%)
CPU Qing (%)
IO Busy (%)
IO Qing (%)
Memory Qing (%)
Network Busy (%)
Projection: 13
Projection Name: 130
llspwd Response Scenario Graph - ROSAGUPG
Projected Response Time for llspwd
llspwd Response (secs)
Projections
Current Model: ROSAGUPG
Current Projection: 0

# SLAM reports

HQ Day Status

Home | HQ Day Index | HQ Day Status | HQ Week Index | HQWeek Status | Top Day Index | Top Day Status

**Status Report**

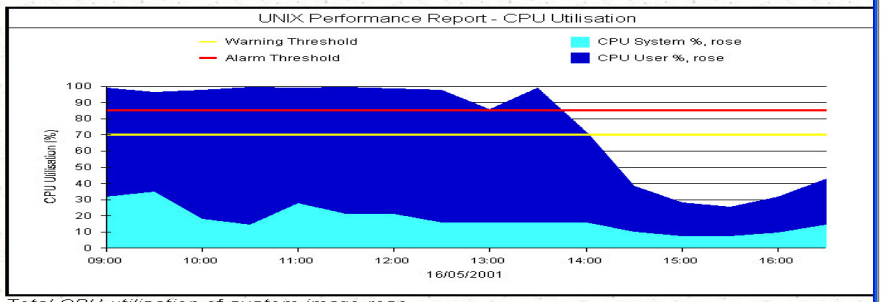
**Bulletin Name:** Head Office servers - Daily Performance
**Installation:** Wkshop73
**Analysis Period:** 16/05/2001 to 16/05/2001
**Date and Time of Report:** 31/10/2001 08:54

**Status Report**

| System | Oracle Database(s) | CPU | Memory | I/O | Oracle | Other |
|---|---|---|---|---|---|---|
| daffodil | N/A | | | | N/A | |
| rose | N/A | | | | N/A | |
| snowdrop | N/A | | | | N/A | |
| tulip | N/A | | | | N/A | |
| crocus | N/A | | | | N/A | |

UNIX Performance Report - CPU Utilisation - Microsoft Internet Explorer provided by PIP...

## UNIX Performance Report - CPU Utilisation




*Total CPU utilisation of system image rose*

### Interpretation of UNIX Performance Report - CPU Utilisation

**Average value compared with thresholds**

The total CPU utilisation, averaged over the analysis period, is 75.8%. This is rat
high, because it is above the lower (warning) threshold of 70.0%, but less than the
(alarm) threshold of 85.0%. There is some, but not very much, spare CPU capacity
handle peaks in the workload without encountering performance problems.

**Proportion of Time compared with thresholds**

The amount of time when the total CPU utilisation is Good is 2.76 hours or 34.4%
total time. The amount of time when the total CPU utilisation is Rather High is 0.5
or 6.3% of the total time. The amount of time when the total CPU utilisation is
Excessively High is 4.77 hours or 59.4% of the total time.

HQWeek Status - Microsoft Internet Explorer provided by PIPEX

HQWeek Status

Home | HQ Day Index | HQ Day Status | HQ Week Index | HQWeek Status | Top Day Index | Top Day Status | Top Week Index | Top Week Status

**Status Report**

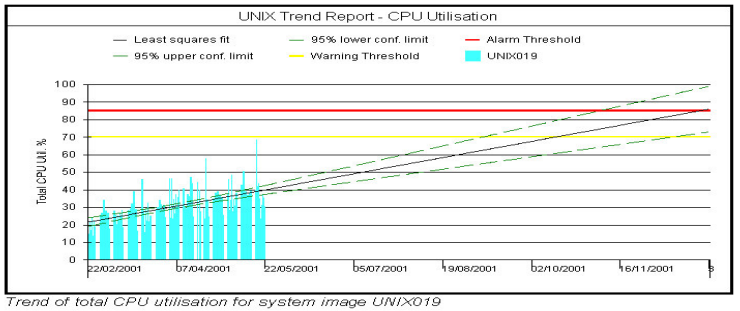
Bulletin Index

**Bulletin Name:** Head Office servers - Weekly Trend
**Installation:** Wkshop73
**Analysis Period:** 13/05/2001 to 19/05/2001
**Date and Time of Report:** 31/10/2001 09:05

**Status Report**

| System | Oracle Database(s) | CPU | Memory | I/O | Oracle | Other |
|---|---|---|---|---|---|---|
| daffodil | N/A | | | | N/A | N/A |
| rose | N/A | | | | N/A | N/A |
| snowdrop | N/A | | | | N/A | N/A |
| tulip | N/A | | | | N/A | N/A |

UNIX Trend Report - CPU Utilisation - Microsoft Internet Explorer provided by PIPEX - [Working Offline]




*Trend of total CPU utilisation for system image UNIX019*

### Interpretation of UNIX Trend Report - CPU Utilisation

**Rate of Change**

The total CPU utilisation is increasing at the rate of 0.2 percent per day. This is equal to 1.4 percent per week, or approximately 6.3 percent per calendar month.

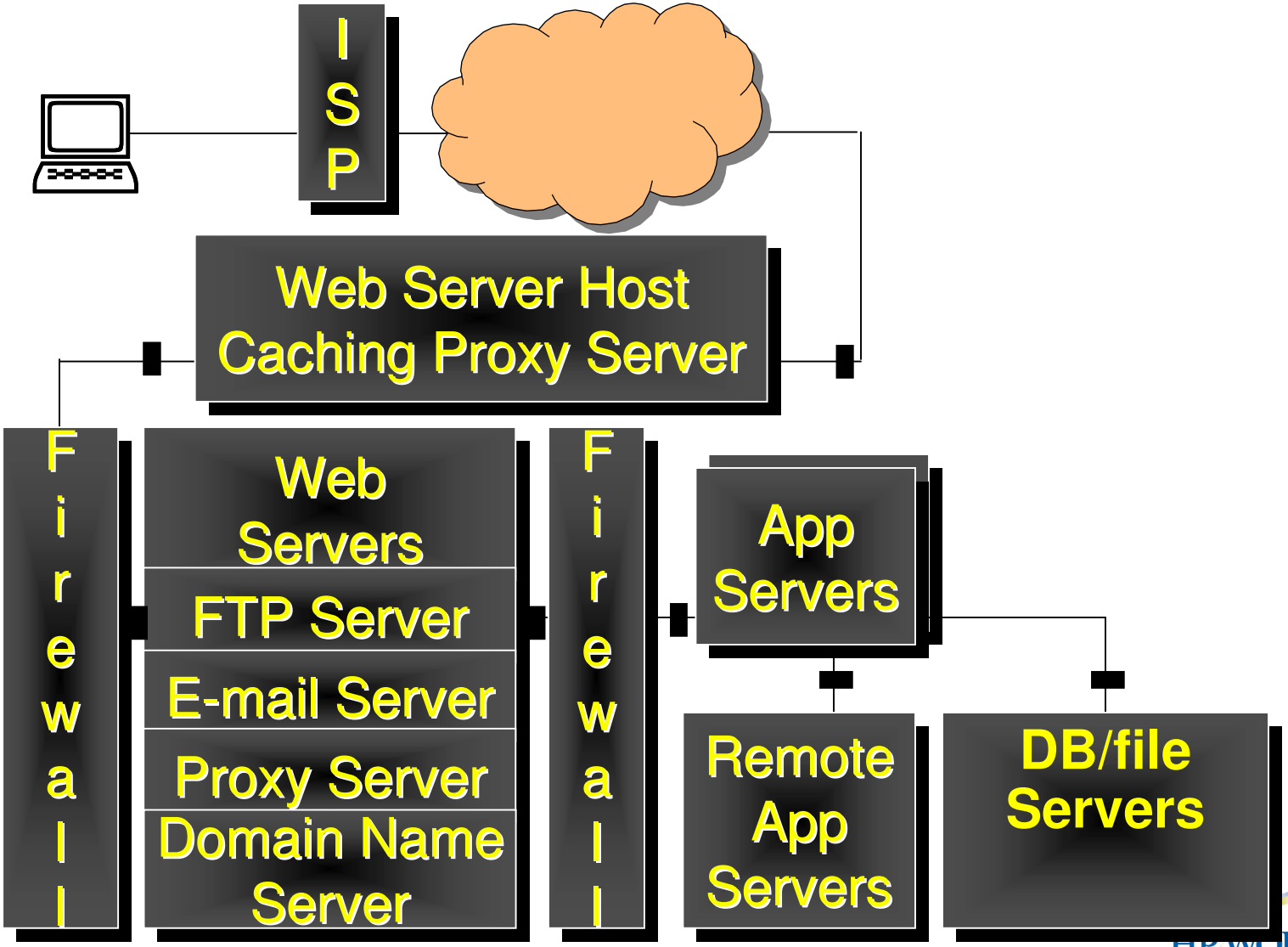
**Value at end of projection**

The projected total CPU utilisation reaches the specified value of 86.0 percent on 30/12/2001.

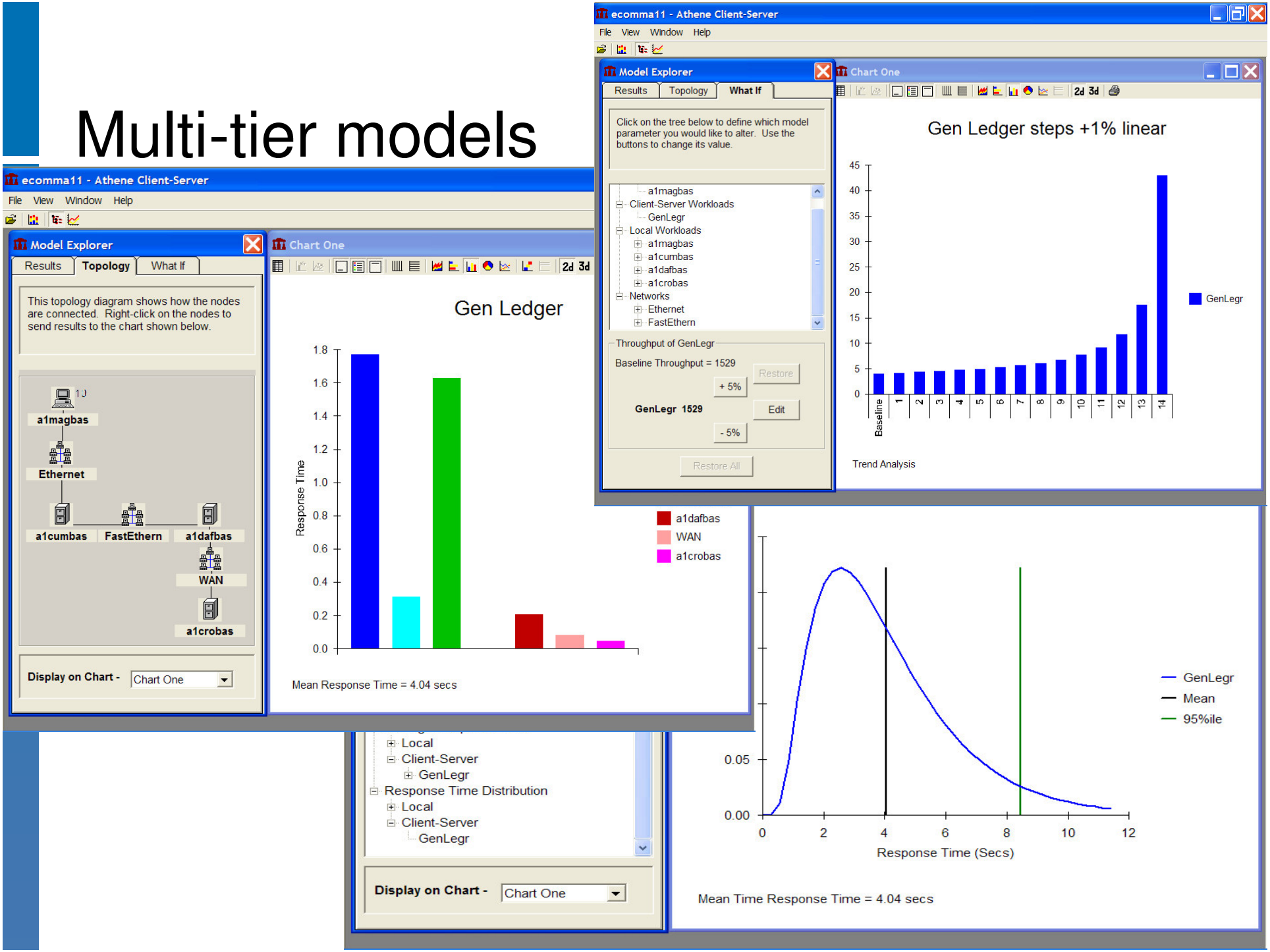
**Projected value compared with thresholds**

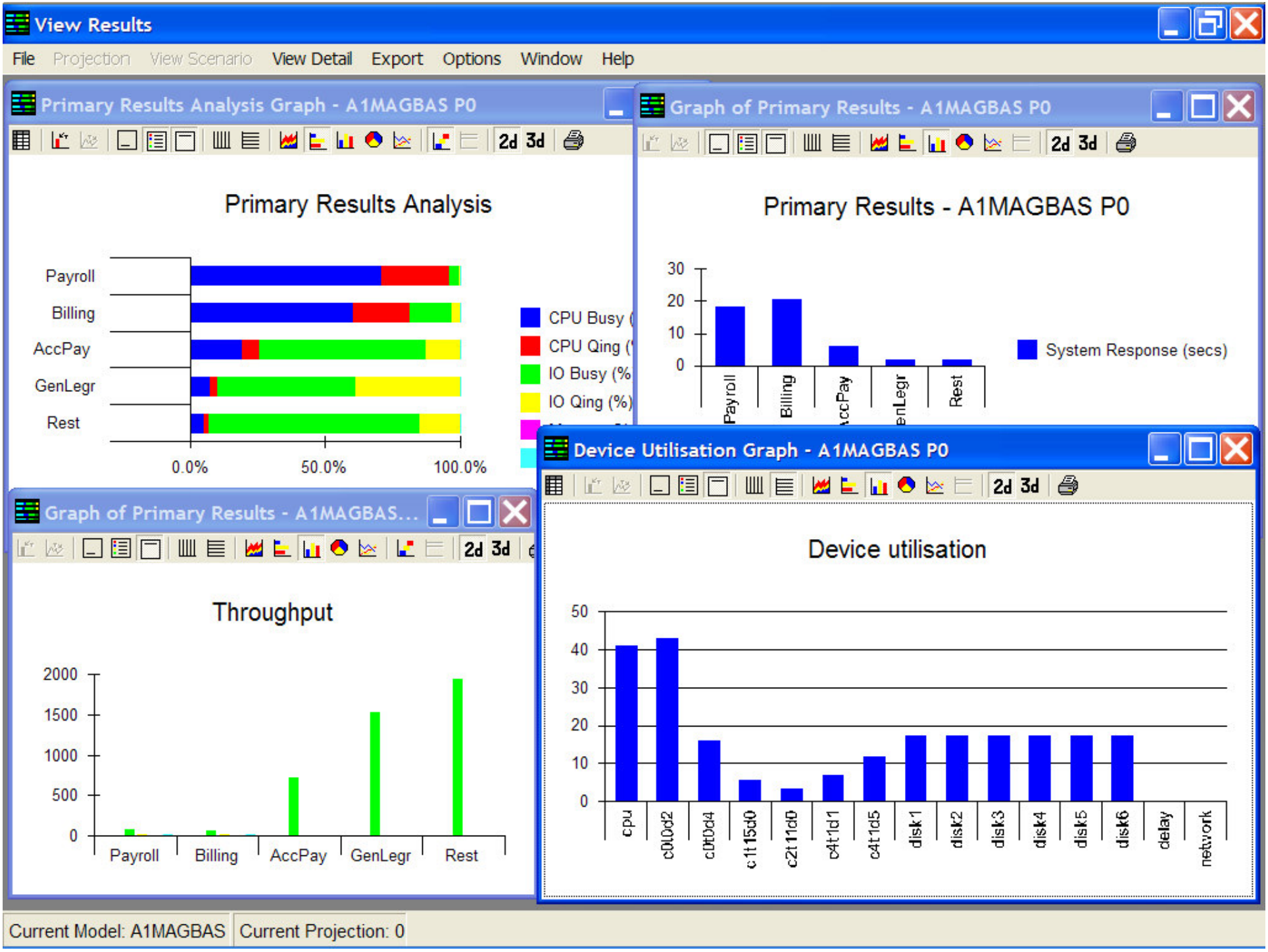
The projected total CPU utilisation rises above the upper (alarm) threshold of 85.0 percent on 25/12/2001.

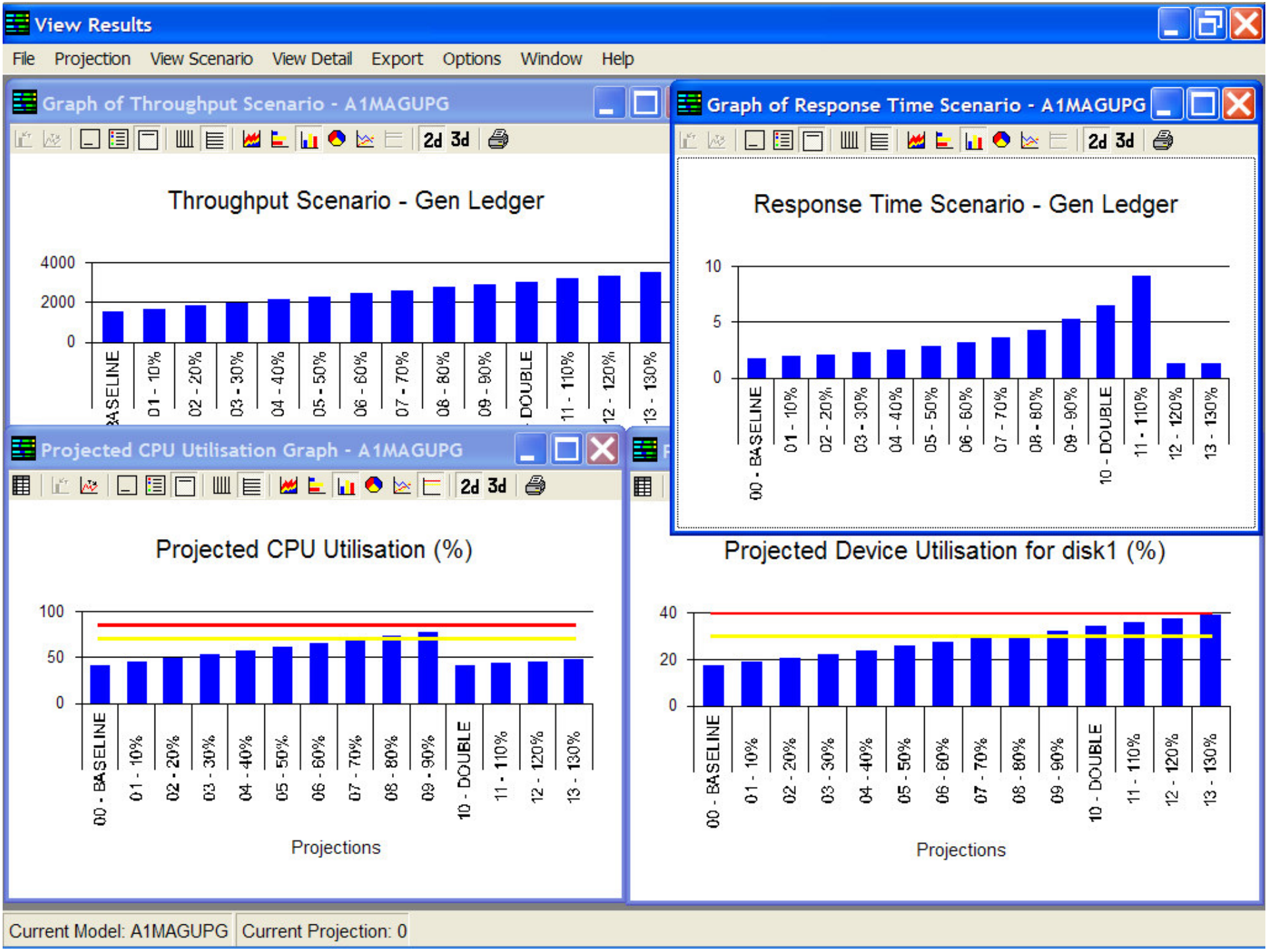
# E-Commerce Multi-Tier Solution



# Multi-tier models




View Results
File
Projection
View Scenario
View Detail
Export
Options
Window
Help
Primary Results Analysis Graph - A1MAGBAS P0
2d 3d
Primary Results Analysis
Payroll
Billing
AccPay
GenLegr
Rest
0.0%
50.0%
100.0%
CPU Busy
CPU Qing
IO Busy (%
IO Qing (%)
Graph of Primary Results - A1MAGBAS P0
2d 3d
Primary Results - A1MAGBAS P0
30
20
10
0
Payroll
Billing
AccPay
GenLegr
Rest
System Response (secs)
Graph of Primary Results - A1MAGBAS...
2d 3d
Throughput
2000
1500
1000
500
0
Payroll
Billing
AccPay
GenLegr
Rest
Device Utilisation Graph - A1MAGBAS P0
2d 3d
Device utilisation
50
40
30
20
10
0
cpu
c0t0d2
c0t0d4
c1t15d0
c2t11d0
c4t1d1
c4t1d5
disk1
disk2
disk3
disk4
disk5
disk6
delay
network
Current Model: A1MAGBAS
Current Projection: 0


View Results
File
Projection
View Scenario
View Detail
Export
Options
Window
Help
Graph of Throughput Scenario - A1MAGUPG
2d 3d
Throughput Scenario - Gen Ledger
4000
2000
0
01 - 10%
02 - 20%
03 - 30%
04 - 40%
05 - 50%
06 - 60%
07 - 70%
08 - 80%
09 - 90%
11 - 110%
12 - 120%
13 - 130%
Graph of Response Time Scenario - A1MAGUPG
2d 3d
Response Time Scenario - Gen Ledger
10
5
0
00 - BASELINE
01 - 10%
02 - 20%
03 - 30%
04 - 40%
05 - 50%
06 - 60%
07 - 70%
08 - 80%
09 - 90%
10 - DOUBLE
11 - 110%
12 - 120%
13 - 130%
Projected CPU Utilisation Graph - A1MAGUPG
2d 3d
Projected CPU Utilisation (%)
100
50
0
00 - BASELINE
01 - 10%
02 - 20%
03 - 30%
04 - 40%
05 - 50%
06 - 60%
07 - 70%
08 - 80%
09 - 90%
10 - DOUBLE
11 - 110%
12 - 120%
13 - 130%
Projections
Projected Device Utilisation for disk1 (%)
40
20
0
00 - BASELINE
01 - 10%
02 - 20%
03 - 30%
04 - 40%
05 - 50%
06 - 60%
07 - 70%
08 - 80%
09 - 90%
10 - DOUBLE
11 - 110%
12 - 120%
13 - 130%
Projections
Current Model: A1MAGUPG
Current Projection: 0


# SLAM DUNK. How to score full points on performance in SLAs.

**Adam Grummitt – adam.grummitt@metron.co.uk**

Director
Metron

Any Questions ???

# Conclusion

- Small overhead to add performance to SLAs
- Without it, there is no performance assurance
- Only a measurable SLA can be used to police
- Modelling enables meaningful measures
- Both sides of the service have an agreed measure
- Performance of service becomes a known entity
- The service level is a sure thing; it's a SLAM dunk!

# HP WORLD 2004

Solutions and Technology Conference & Expo

Co-produced by:

interex

shared knowledge • shared power

encompass

AN HP USER GROUP