SIEMENS

Best Practices for Writing and Managing Performance Requirements

Andre B. Bondi
Siemens Corporation
Corporate Research and Technology
Princeton, New Jersey 08540 USA
andre.bondi@siemens.com

SIEMENS

Overview of the Problem

SIEMENS

Background

- Performance problems are a major cause of project failure. They tend to have their roots in
  - inadequately drafted performance requirements

  and
  - poor architectural and design decisions
- Performance, reliability, and availability considerations must be built into the system from the time it is conceived.
- ➢ Consistent and unambiguous requirements specifying performance, reliability, and availability are preconditions for a successful system.

SIEMENS

Motivation I: Problem Prevention

- Poorly written performance requirements cause confusion.
- A badly written performance requirement could be unattainable.
- Performance requirements often turn up in legal contracts.
- Well written quantitative requirements lead to a better product that is delivered sooner with less effort.

❖***Direct impact on cost and profitability of the product.***

SIEMENS

Motivation II: Driver of Architectural Decisions

- Performance requirements have a fundamental impact onf the computer system architecture
  - ❑Throughput and response time sustainable by proposed technology?
  - ❑Focus of overload inherent in the architecture or design?
  - ❑Can the architecture exploit parallelism if desired?
  - ❑Is the architecture cost-effective?
  - ❑Are the performance requirements achievable?

SIEMENS

Motivation III: Defining Performance Tests

- Performance requirements should form the basis for meaningful, coherent performance tests
  - Tests needed to determine if the performance goals of the system have been attained.
  - The requirements spell out the goals.
- Without performance requirements, performance testing is prospective, unfocused
  - Results have no or inadequate context.

SIEMENS

Confusion About Performance Requirements is Costly!

- It hurts developers' morale and delays delivery of the system.
- Stakeholders (including architects, developers, and testers) must spend time together deciding what a poorly written quantitative requirement actually means.
- The performance of the system on delivery may be unacceptable.
  - ***Unhappy users, and/or useless system!***
- Poorly written requirements are often included in contracts.
  - ***Financial penalties, legal disputes!***

SIEMENS

Possible Causes of Confusion

- Inadequate and/or incorrect specification of requirements that are essentially quantitative
- Quantitative requirements that are mutually inconsistent, that are in conflict.
- Slogans masquerading as requirements
- No apparent **link** to business, engineering, or regulatory needs, or to the problem domain.
- Requirements written in terms that are **not measurable**.
- Etc.

SIEMENS

Examples of Vague and Specific Performance Requirements

- "The system shall be scalable."
- "The system shall be up all the time."
- "The system shall support all transactions submitted to it."
- "The response time shall be less than 2 seconds 99 per cent of the time."

- *A small system shall host 100 users, a large one 1,000 users.*
- *The system shall be available 99.99% of the time.*
- *In the peak hour, the transaction rate is 10/sec for a small system, 1,000/sec for a large one.*
- *In the peak hour, the average response time shall be 2 seconds or less, and 95% of the transactions shall have a response time of 5 seconds or less.*

SIEMENS

Need Definitions of Performance

- Need working definitions of system performance, before writing requirements about it.
- Need *metrics* to describe it.
- Requirements must be unambiguous and testable to be enforceable.

SIEMENS

What is System Performance? What Metrics Specify It?

SIEMENS

Describing System Performance

The description depends on
- the problem domain,
- what the system does,
- how it is used,
- the other systems with which the system interacts,
- how it generates revenue,
- whether it is mission-critical,
- the causes of (external) load, the sources of requests for actions.

SIEMENS

"User Experience" Performance Metrics

- Throughput (load) or capacity
- Response time, waiting time *for specific tasks*
- Number of "entities" supported concurrently (average, max)
- Transaction success rate or failure rate
- Message delivery and loss rates
- Transaction success rate
- Transaction failure rate

o Some of these metrics can be used to specify service level agreements (SLAs).

SIEMENS

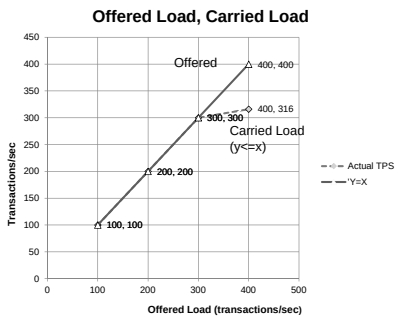
Performance Metrics and Resource Utilization Metrics

- Performance metrics can be affected by, and are related to, *resource usage metrics,* for example:
  - bandwidth utilization,
  - CPU utilization,
  - memory occupancy ("average number of pages"),
  - disk utilization.
- Resource utilization is an indicator of carried load, but *not* a direct measure of performance.
  - High utilization => resource close to saturation
  - Low utilization = { resource has spare capacity } OR { path to it is obstructed }

SIEMENS

Offered Load, Carried Load

- *Offered load* is the load presented to the system
- *Carried load* is the load that is actually carried, that gets through, sometimes called *goodput* or *throughput*.



SIEMENS

Describing Performance in Different Domains

Requirements and Experience

1. The average response time of an ATM is 2 seconds.
2. The average time from a mouse click on a link to the complete display of the next web page is 4 seconds.
3. The conveyor system can handle 1,000 suitcases per hour.
4. "I had to wait 35 minutes to get through security!"
5. The airbag should be deployed within *x* milliseconds of impact.
6. The surveillance system shall need no more than 10 seconds to process and display a burst of 100 alarm message arriving within 5 seconds.

Questions to Ask

1. *How many other customers were using ATMs at this bank when you were?*
2. *What is delivered in response to the click? Images? Responses to database query?*
3. *How long did it take to deliver them?*
4. *Measured from where to where? The time you saw the first identity screener? The time you joined the queue for the first screener?*
5. *The airbag is mission critical. Does it have a dedicated controller? Are there performance requirements on that controller?*
6. *Why is the average alarm handling rate not a helpful metric here?*

SIEMENS

More on Performance Metrics

Performance metrics should inform us about
- The domain of the system
- The capacity of the system
- The demands made on the system
- The quality of service provided by the system

Performance requirements should be expressed in terms of metrics that are
- informative about the domain
- measurable
- expressed in the correct time scale (per second or per month?)

SIEMENS

Beware of Mononumerosis!

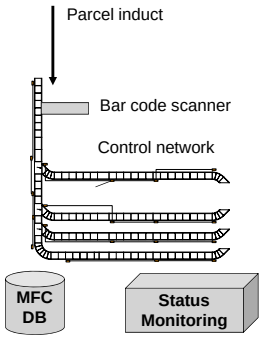
Mononumerosis *n.*: 1. a tendency to fixate on a single number (A. M. Odlyszko, CMG Magazine, 2000). 2. an undue focus on a single measured value.
- One cannot rely on a single number to tell the whole story.
- A description of the performance requirements of a system requires *context,* such as the offered load (job arrival rate, what is to be done) and the desired response time.

➢A performance requirement based on an ill-chosen single number is insufficiently specific to tell us how well the system should perform.

SIEMENS

Example: Conveyor System

Domain-related metrics:
- Parcels per second at each point in the conveyor
- Conveyor speed (metres/sec)
- Distance separating parcels
- MFC DB (parcel routing database) response time: must know which way a parcel is to be diverted before it arrives there.
  - *What drives the query rate?*
- Message rates and latencies between programmable automation controllers
- Volume of status monitoring traffic
- Status notification time: time from status change to notification on display
  - Conveyor malfunction
  - Emergency cord pulled => line shut down within X seconds



SIEMENS

Example: Train Signaling and Departure Boards

As trains move from station to station,
- Display how long it will be until a train to a particular destination arrives at each station.
  - Display is on indicator boards or monitors on each platform and at access point.
  - E.g., at Columbus Circle on the A and D lines, next downtown A train 3 minutes, next downtown D train 2 minutes.
  - E.g., at Copley, next to Heath E in 4 minutes, next to Riverside in 8 minutes.
- Indicate position of each train in a central signal box or control room
- Keep trains safe distances apart
- In case of emergency, trigger stop signals in some zones but not others, and notify passengers waiting on the platforms

SIEMENS

Metrics for Train Signaling and Departure Boards

Performance Objectives

- Time from train movement to completion of updates of platform indicator boards. *Information displayed at each station is different. Why?*
- Time from train movement to completion of control room updates
- Time from train movement to turn signal red in block immediately behind it, and green two blocks behind it.

Load and Requirement Drivers

- Number of train movements per minute
- Speed of each train
- Average time from arrival to departure of each train at each station. Also, how long may the doors be open?
- Number of trains en route or at intermediate stations
- Number of stations to be notified about each train
- Travel time between stations

SIEMENS

Example: The Fitness Center

- Critical resources:
  - Lockers, stationary bikes, showers, various kinds of exercise machines,….
- Metrics:
- Requirements: (???)
  - No queueing for equipment
  - Lockers always available
  - Showers always available

SIEMENS

Example: Radiology Processing and Display

Example Functionalities

1. Image retrieval
2. Image annotation
3. Image movement (rotation, zoom, etc.)
4. Image storage and preprocessing

Performance Metrics

1. Response time, number of times per hour
2. Response time
3. Response time (depends on the movement)
4. Storage time, preprocessing time

SIEMENS

Example: Fire Alarm System

Example Functionalities

1. Sounding an alarm
2. Playing evacuation announcements
3. Closing vents and doors
4. Displaying problem location information
5. Notification of malfunction on control panel

Performance Metrics

1. *Time from detection to sounding alarm; time from pulling red handle to alarm*
2. *Time from detection to commencement of play*
3. *Time from detection to action*
4. *Time to handle alarms from N sources*
5. *Time from detection to notification*

SIEMENS

Metrics

Some Definitions, Characteristics, and Examples

What is a metric?

Definition: A metric is a standard of measurement.
---Webster.

- Example:
  - Average response time of a transaction *of a given type*
- Counterexample:
  - *"There is no metric for happiness."*

Notes:

- A metric can be derived from one or more measurements.
  *Utilization = Throughput * (Average Service Time)*
- A metric may be obtainable in theory but not in practice.

Definitions of Some Metrics

- The response time is the elapsed time between the initiation of a transaction and its completion.
- The average response time is a *sample* statistic collected during a specific time interval, e.g., if $N(T)$ tasks completed in [0,T]:

$$\overline{R} = \frac{\sum_{i=1}^{N(T)} R_i}{N(T)}$$

- The average utilization and average queue length are *time-averaged* statistics.
- For a single resource,
  $U(t)$ = 1 if the resource is in use at time t,
  $U(t)$ = 0 otherwise.
- Average utilization of a single resource during [0,T]:

$$\overline{U} = \frac{\int_0^T U(t)\,dt}{T}$$

Example of an Unobtainable Metric

$$S^2_{N-1} = \frac{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}{N-1}$$

Desired: variance or standard deviation

- System collects sum of the observations, average, min, max, sample size
- System does not collect sum of squares, does not save individual observations

➢ Cannot estimate the variance or standard deviation

SIEMENS

Useful Characteristics of a Metric

- ❖ Reference: Lilja, *Measuring Computer Performance,* Cambridge University Press, 2000.
- Linearity. If the metric changes by a certain ratio, so should the performance it quantifies.
- Achievable throughput characterizes performance.
- Reliability. A metric is considered *reliable* if System A outperforms System B whenever the metric says it does.

SIEMENS

Useful Characteristics of a Metric II

- *Example*: system response time
  - ✓ Reliable
  - ✓ Linear
  - NB: user-experience metric!
- *Counterexample.* MIPS is an unreliable metric
  - Some programs could take longer to execute on a processor with more MIPS than one with fewer, or vice-versa.
  - Possible reasons: instruction sets have different functionalities, cache memories might perform differently, different I/O mechanism, etc.

SIEMENS

Useful Characteristics of a Metric III

- Repeatability. One obtains the same value of the metric every time one repeats an experiment under identical conditions => Metric is *deterministic.*
- Ease of measurement.
- Consistency. The definition of the metric and the units in which it is expressed are the same across all systems and configurations.
  - ➢ This is not always true of MFLOPS or MIPS.
- Independence. The metric should not reflect the biases of any party or stakeholder, or it will not be trusted.
  - Reason: A stakeholder might choose to base comparisons on a metric that is favourable to his system.
  - E.g.: MIPS vs. execution time of a known benchmark.

SIEMENS

Choice of metric can depend on who will use it, for what purpose.

Example:

- Number of transactions per month.
  - ❑ *Business-related metric.*
  - ❑ Of interest to CFO, accountants, etc.
- Number of transactions in the busy hour. *Engineering-related*, *service-related metric.*
  - ❑ Necessary for correct sizing of the system to ensure good response time etc.
  - ❑ Of interest to capacity planners, systems administrators, etc.

➢ The two metrics are related through (possibly unstated) assumptions about *or knowledge of* traffic patterns.

➢ Neither metric is useful or meaningful to the other stakeholder by itself.

SIEMENS

Metrics Related to User Experience

- Response time
  - Time from initiation of a transaction or action to its completion.
  - Web:
    - Sometimes defined as the time from a click to the appearance of the first byte of the response
    - Sometimes defined as the time from a click to the complete delivery of the requested page.
    - *Ask!*
  - At airport security, time from joining the line to leaving the checking area (*before* you put your shoes back on. Why?)
- Waiting time
  = Response time - service time
  = Time to get to the head of the line
  - At security, does not include the time to X-ray luggage or search passengers.

SIEMENS

Revenue-Related Metrics

- Arrival rate (charge per offered job = potential revenue)
- Throughput (charge per handled job = actual revenue)
  - Sometimes called the completion rate
- Blocking probability (probability call not accepted)
- Balking probability (customer goes away of own accord)
- Lost calls rate = (Call arrival rate) × (Blocking probability)
- Abandonment fraction (calls dropped by customers balking or customers leaving queue if they do not wish to wait any longer)
- Buffer overflow probability (probability packets lost because buffer is full)

SIEMENS

Metrics Indicating Customer Dissatisfaction

- Forced call drop rate ("Please try later…")
- Abandonment rate
- Retry rate

Note: Retries may be desirable from a revenue standpoint but not a cost standpoint. Why?

SIEMENS

Some System Resource Usage Metrics

- Server utilization
  - CPU busy, I/O busy, average number of servers busy, bandwidth utilization.
    - Parallel processors: utilization of *each* CPU is of interest. An imbalance indicates a possible problem.
- Occupancy
  - Current, average number of [memory pages, hospital beds, etc.] in use
  - Current, average number of occupied chairs in barber's waiting area
  - Average queue length, max queue length
  - Swap space size
- Page fault rates (different types…)

SIEMENS

Some Network-Related Metrics

- Offered packet rate (packets/second)
- Carried packet rate (packets/second)
- Packet loss probability =
  - 1 – ((carried packet rate)/(offered packet rate))
- Bit error rate
- Average, min, and max packet sizes (bytes)

SIEMENS

Measures of System Demand

Explicit:

- Average rate for each transaction type
- Average service time per transaction for each transaction type

Implicit:

- Number of users logged in
  - ❖ demand per user must be specified for this metric to be unambiguous

SIEMENS

Beware of Seasonal Variation

- Demand varies by time of day, season, day of month, minutes past the hour, etc.
- Important to know the peak demand so that the system can be engineered to support good performance all the time, as well as handling traffic spikes.
  - If not all the time, at least determine how bad the performance can be without incurring too much lost revenue or ill will.
  - Performance problems typically occur at the peak time.

SIEMENS

Know Sources of Measurements!

- ❖ If you cannot measure the quantities listed in a performance requirement, you cannot enforce it.
  - Legal implications
  - Must be able to measure performance characteristics of any component to be integrated into a larger system.
- ❖ You must know the sources of your measurements when writing the requirements.

SIEMENS

Possible Sources of Measurements

- System resource usage:
  - Operating system measurements and those derived from them
- Transaction volume:
  - The application itself
  - In web-based systems: access logs
  - In database systems: transaction logs, journals
  - Packet traces, by inspection of the payload fields and filtering
- Response Times:
  - Load generators used as probes (dummy transactions)

SIEMENS

More About "Numbers of Users"

- Requirement 1: "The system shall host H users."
- Requirement 2: "The system shall support a maximum of N and an average of A concurrently logged in users."

Why are these requirements incomplete?

- We do not know what "support" means.
- We do not know the "footprint" required by an active user, nor that of an inactive user.
- We do not know often a user logs in and how long he stays logged in, or at what time of day.
- We don't know what users do when logged in, or how often, or the processing requirements of their activities.

➢Always need to specify the demands made by a ***reference user*** before specifying anything about numbers of users.

SIEMENS

Must State Basic Traffic Assumptions to Set Performance Context!

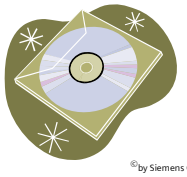
By stating traffic assumptions,

- We reduce the risk of ambiguity
- We obtain a better understanding of the system load

➢Need to define reference user:

- Executes a defined set of tasks *x* times per second
- Has a footprint of *M* kB memory and uses *D* kB of disc space

➢Need to define a reference workload.



SIEMENS

Exercise 1

- Why is the number of users logged in NOT a sufficient, consistent, or reliable indicator of user demand?
  - What other metrics and information do we need to complete the picture of user demand on the system?
- _____
- _____
- _____

SIEMENS

Performance Requirements and Data are Commercially Sensitive!

- The performance of a product or system can be a differentiator in the market place.
  - Performance data can tell us a lot about the limitations or potential of a product or system.
- Knowledge of performance data can enable one to figure out a competitor's volume of business, or the maximum volume that can be attained. One can then put the competitor at a disadvantage.

➢Performance requirements and data should be treated as confidential until cleared for release.

SIEMENS

Exercise 2: Performance of the Refreshment Service in Class Breaks

- Identify the constraining factors.
- Identify the demand variables.
- Identify criteria for satisfactory performance.

➢Identify metrics

- describing the work done
- describing the performance of the system

➢Explain which objectives are served by

- having all tutorials on break at the same time,
- having tutorials on break at different times.

SIEMENS

Writing Performance Requirements

SIEMENS

Performance Requirements and Functional Requirements

Like functional requirements, performance requirements should be

- Traceable: why is each one there?
- Based on a well defined and specified set of assumptions.
- Verifiable: is each one correct?
- Measurable:
  - Is there a metric describing it?
  - Can we obtain the metric? How? From what source?
- Modifiable: what effect will changing a performance requirement have on the rest of the system? On other requirements?
- Unambiguous: each should only have one meaning.

❖Terminology: IEEE Std 830-1998

SIEMENS

General Guidelines for Writing performance requirements

- The statement of each requirement should be separated from its supporting commentary
- Performance requirements must be described by sound metrics and be physically achievable and testable.
- Performance requirements must meet business and engineering needs.
  - Business needs should be linked to regulatory requirements and income streams.
- Demands made by reference users should be well defined so that the implications of requirements expressed in terms of the number of users can be easily understood.
- Circular dependence between requirements should be avoided.
- Requirements should be parameterized when the values of quantities on which they depend are not yet known.

SIEMENS

Necessary Attributes of performance requirements

- Each should be written in terms that are measurable, verifiable, and linked to engineering and business needs.
- Each should be *traceable*.
- They should be consistent with one another, algebraically and functionally

SIEMENS

Example

- The system shall sustain a throughput of 1 transaction per second.
- The system shall support 100 users, each of whom generates 0.1 transactions per second.

Problems

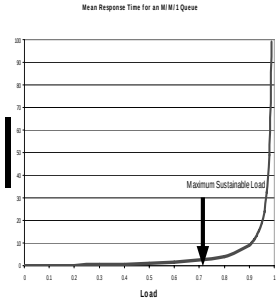
- Requirements 1 and 2 are inconsistent
  - Requirement 2 => 0.1x100 = 10 transactions/second != 1 transaction/second.
- The terms *sustain* and *support* are ambiguous in the absence of definitions of their meaning.

SIEMENS

Working Definition of *Sustain:*

A specified transaction mix can be *sustained* at a given rate if <u>all</u> of the following hold:

- The requirements for the average response time and the probability that a response time exceeds a certain level are both met
- A small increase in the transaction rate will cause only a small increase in the average response time.

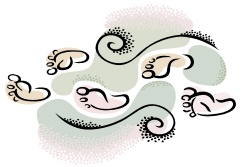


SIEMENS

Working Definition of Supporting N Users

We can support *N* users if <u>all</u> of the following hold:

- The combined load they generate is sustainable by all resources (CPU, I/O, network, etc.)
- Their concurrent footprints fit into memory without causing so much paging that the load would not be sustainable.
- There is enough room for their combined disc footprint.



SIEMENS

Questions To Be Answered by Traceability

- How was the value of each quantity chosen or computed? What formulae were used? Why?
- Why was each metric used in the performance requirement?
- What business needs are met by specific values in performance requirements? *E.g.: throughput to meet a customer need.*
- What engineering need is met by specific values in performance requirements? *E.g.: protocol timeout drives choice of response time.*
- Who requested this requirement? Why?
- Who wrote this requirement?

SIEMENS

Example: Performance Requirement

Wrong:

- Ideally, the response time shall be at least one second.
- The response time shall be at most 2 seconds.

❖*If a single response time exceeds 2 seconds, the requirement will not be met. But this COULD happen, even if it is unlikely. It is better not to specify the requirement this way.*

Correct:

- *The <u>average</u> response time during the busy hour shall be at most 1 second.*
- *99% of all response times shall be less than 2 seconds during the busy hour*.
- *Both requirements shall be met simultaneously.*
- Note: NOT the same as saying that the response time will be less than 2 seconds 99% of the time. *During what time period? At what load?*

SIEMENS

Functional Requirements vs. *Performance Requirements*

- A *functional requirement* specifies what a system shall do or produce with a given set of inputs.
- One can test that a functional requirement is met for a given input by checking that the system produces the expected output.
- A functional requirement must be verifiable and unambiguous.

- A *performance requirement specifies temporal characteristics of the system, e.g.,*
  - *the rate at which a system shall handle units of work,*
  - *how many work activities must be supported concurrently,*
  - *how long it takes to handle or produce a unit of work.*
- *To be unambiguous, a performance requirement must be posed in <u>measurable terms</u>.*
- *To be useful, a performance requirement must be specified with a level of time granularity commensurate with the time scales for which the system must be engineered.*

SIEMENS

Criteria for Quantitative Requirements

- The requirement should be written in terms that are unambiguous, measurable, and verifiable.
- The metrics used to formulate the requirement should have direct impacts on service and safety.
- Every requirement must be traceable.
  - Who wrote it, and why? Reference?
  - Mathematics used to determine quantities?
  - Where did the parameters come from?
- Requirements must be mutually consistent, algebraically and functionally.
- Every requirement must be linked to a service, engineering, business, operational, regulatory, or safety need.

SIEMENS

Example: The Fitness Center

- Critical resources:
  - Lockers, showers, various kinds of exercise machines,….
- Metrics:
- Performance Requirements: (???)
  - No queueing for equipment
  - Lockers always available
  - Showers always available

❖*What is wrong with these requirements?*

SIEMENS

Examples of Assumptions About Metrics

Metric: calls per hour.

- *Insufficient:* A "typical" telephone call is assumed to last 3 minutes (industry convention). **What services are involved in the call? *Prepaid card, credit card, voice mail, 800 routing…?***

Metric: number of concurrent users logged into a web site.

- *Insufficient:* one should also specify what these users are assumed to be doing and how often; the average memory footprint per user, the amount of back end and server processing per click, etc.

Metric: A web site is expected to handle X user sessions per month.

- Metric is OK for forecasting revenue.
- Metric is *insufficiently informative* about the busy hour volume.

SIEMENS

Why We Need Unambiguous Performance Requirements Expressed in Measurable Terms

- Performance requirements may be embedded in a contract between a buyer and a seller. Unless they are expressed in measurable terms, it may be very difficult to enforce them.
- The revenue stream generated by the system often depends on the ability to meet performance requirements.
- The system cannot be sized economically unless the user's performance demands are understood.
  - Therefore, one must also check that a customer's requirements are consistent with market conditions.
- There are social and economic consequences for failing to meet performance requirements.
- If a control system fails to meet a performance requirement, or if the requirement is ill posed, system malfunctions may occur, perhaps with catastrophic results.
- Poorly drafted requirements cause systems engineers and developers to spend a lot of time in unproductive meetings, or in court. This delays delivery of the product and increases costs.

SIEMENS

What do we mean by "measurable terms?"

- There has to be a means for time-stamping, measuring and logging values of the quantities of interest to an appropriate level of granularity (per sec, per minute, per hour, etc.).
  - The logs must be readable after the fact.
- The measurements may be collected by the operating system, by outside instrumentation, or by hooks embedded in the applications of interest.

SIEMENS

When negotiating performance requirements…

- Ensure that every quantity about which there is a requirement is *measurable* and/or *derivable* from direct measurement.
- Ensure that the performance requirements are consistent (this may require doing some math!).
- Make sure that there is a good business or engineering reason for every performance requirement (traceability).
- There must be a good engineering or business reason for collecting every measurement, or you will be drowned in data.
- Make sure that the lawyers and "suits" negotiating an agreement understand the need for measurability, so that they don't create unverifiable performance requirements. [This means there have to be good links between Marketing, Sales, and Engineering.]

SIEMENS

External Performance Requirements and Derived Performance Requirements

- If we specify that the peak system transaction rate is X per second, and each transaction uses device *j Vj* times per second, device *j* must be able to support a load of *XVj* transactions per second.
- If the transaction response time is *R,* the time spent at device *j* is *VjRj.* Hence, we require

$$Rj < R/Vj \text{ for all } j.$$

❑The desired values of *X* and *R* are *external* requirements.

❑The performance requirements on device *j* that follow from the external requirements are derived requirements.

❖Ensure that the stakeholders know: *IMPROVES TRACEABILITY!*

SIEMENS

Exercise 3

- You are told that your corporate system has to support 5,000,000 transactions per month. The total number of potential users is 50,000. The application can only be accessed inside the firewall. Identify other factors and/or assumptions that are needed to specify the load. *Hint: will all users log on at once?*
- _____
- _____
- _____

SIEMENS

Guidelines for Writing Performance Requirements

Structure of a Performance Requirements Document

Outline:

1. Scope and Purpose
2. Intended Audience
3. References (including functional requirements spec)
4. Statement of Assumptions:
   a. Traffic assumptions
   b. Criteria for load sustainability
   c. Definitions of availability and reliability for this system (if applicable)
   d. Definitions of metrics used in the requirements
   e. Instrumentation to gather the metrics for verification
5. Performance Requirements

Needed: Statement of Basic Assumptions

- Document should contain a section stating the basic assumptions underlying everything else within it
  - Explicit assumptions should be stated.
  - Implicit assumptions should also be stated, because they may not be implicit to all.
  - Industry regulations, norms, and standards relating to performance should be explicitly cited and stated.

Examples of Basic Assumptions

- Definition of workload sustainability.
  - This may depend on the problem domain.
- Definition of other terms:
  - *What constitutes an average phone call?*
  - *What is an average transaction?*
- Assumed activities by an average user
  - What uses cases are invoked?
  - How often are they invoked per session?
- Expected reliability and availability

SIEMENS

Examples of Regulations to be Cited

Depends on the domain:

- For fire alarm systems: fire protection code in the applicable jurisdiction
- For railroads: applicable government regulations
- For conveyor systems: government rules governing the function of the emergency pull

SIEMENS

Needed: Description of Metrics Used and Instrumentation to Capture Them

- How is each metric defined? What does it mean?
- Why is each metric being used in a requirement?
- How is each metric linked to a business need? Which need?
- How is each metric linked to an engineering need? Which need?
- Where do the measurements for each metric come from?
  - What instrumentation is needed to capture the measurements?

SIEMENS

Suggested Fields of a Performance Requirement Record

- Requirement Number
- Title
- Statement of requirement
- Supporting commentary
- List of precedents, sources, standards docs (needed for traceability)
- Derivation of quantities
- List of dependent requirements
- List of assumptions and precedent performance requirements, i.e. the performance requirements on which this one depends.
- Sources of measurement data.
- Name of a subject matter expert on this requirement
- Indicator if the requirement is independently modifiable
- Indicator that the requirement is traceable.
- Indicator that the requirement is unambiguous, or if not, why not.
- Indicator that the requirement is correct, or, if not, why not.
- Indicator that the requirement is complete, and if not, why not.
- Indicator that the requirement has passed or failed review, and if not, why not.

SIEMENS

Notes on the Requirement Fields

- The reasons for the requirement belong in the supporting commentary, *not in the statement of the requirement.*
- The chosen values of quantities must be supported by at least one of
  - A reference to a standard or to another requirement
  - A mathematical derivation, or a reference to the formula used to compute it, together with a list of the input parameters and how and why they were chosen.
- The sources of the measurements must be identified or clearly marked as unknown.

SIEMENS

Each Requirement Must Meet Business and Engineering Needs

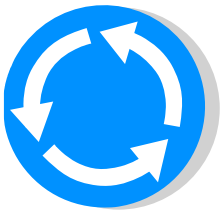
- Throughput is usually related to a business need, such as revenue or estimated demand.
- Response time requirements could be driven by competitive differentiation and/or an engineering need (e.g., avoiding timeouts and retransmissions).
- Business needs may be linked to regulatory requirements.
- ***Business needs should NOT be confused with slogans whose meaning cannot be tested in the lab.***

❖ ***"Always up! Always responsive! Always reliable!" ???***

SIEMENS

Avoid Circular Dependencies

- Circular dependencies make modification of requirements very difficult, if not impossible.



SIEMENS

Some Performance Requirements Pitfalls

SIEMENS

Example: Migration from a Legacy to a Modern System: Metrics Don't Change, But Performance Requirements Could

- Metric of Interest:
  - *Shutter Reaction Time*
  - Elapsed time from when button pushed to image capture

BUT:

- Functionalities are different!
- Shutter reaction is much slower with the new camera
- Overlooked the shutter reaction time when buying the new camera!

SIEMENS

Old Camera vs. New Camera

- Old Camera:
- Setup {
  - •Advance film
  - •Choose aperture, shutter speed from light meter and set on camera
  - •Focus through range finder
  - }
- Press button {
  - /* start shutter reaction time */
  - •Move shutter curtain at correct speed to capture image
  - } /* end shutter reaction time */

- *New Camera:*

Press button lightly {
- o*Autofocus*

}
Press button firmly {
/ start shutter reaction time */*
•Choose shutter speed, aperture
•Autofocus if not done already (??)
•Image processing (??)
•Capture image
} / end shutter reaction time */*

SIEMENS

Pitfall: Performance requirement depends on technique and technology

- Metrics of interest relates to capture of action shots
- With old camera,
  - Can reuse manual shutter, aperture, setting and focus form multiple shots if lighting conditions do not change
  - *Setup and film advance not included in shutter reaction time*
- *With new camera,*
  - *Shutter speed, aperture, and focus are redone every time.*
  - *This affects the shutter reaction time.*

Question:
- Did the choice of metric lead to a fair comparison?
  - o YES, because one desires rapid reaction for rapid sequences of shots
  - o NO, because setup and technique were not included in the comparison

SIEMENS

Topic: Concurrent Programming *Freedom from Deadlock is a Performance Requirement!*

- Reason:
  - If the system goes into deadlock, the throughput goes to zero!
  - This is often overlooked in both the requirements and the design.

SIEMENS

Example: A System Prone to Deadlock on Overload

SIEMENS

Structural Features

- Common FCFS queue for those leaving and collecting coats
- Hangers are non-preemptible, non-sharable resources

Pitfall: potential for deadlock is easily overlooked, especially at light loads.

Exercises:

- How does deadlock occur in this system?
- How can it be resolved *in this system* if it occurs?
- How can it be prevented? Are major architectural changes necessary?
- Will increasing the number of hangers and/or attendants solve the problem?

Next step: photograph the same place when it is cold outside.

SIEMENS

Museum Checkroom in November

SIEMENS

Common Patterns for Performance NFRs

SIEMENS

Common Patterns for Performance Requirements

- Many performance requirements conform to one of a small set of patterns.
- The mistakes they contain also conform to a set of erroneous patterns.
- Examples will show how performance requirements should be drafted.

SIEMENS

Pattern: Required Response Times

Wrong:

•Ideally, the response time shall be at most one second.

•The response time shall be at most 2 seconds.

❖*If a single response time exceeds 2 seconds, the requirement will not be met. But this COULD happen, even if it is unlikely. It is better not to specify the requirement this way.*

Correct:

•*The <u>average</u> response time during the busy hour shall be at most 1 second.*

•*99% of all response times shall be less than 2 seconds during the busy hour.*

•*Both requirements shall be met simultaneously.*

Note: NOT the same as saying that the response time will be less than 2 seconds 99% of the time. *During what time period? At what load?*

SIEMENS

Pattern: Number of Users to Be Supported

Requirement:

- "The system shall support N users."

Problems:

- There is no statement about what the users do, or how often they do it.
- There is no statement about how many users are logged on at the same time.
- There is no distinction between types of users.

Solution:

- Need specifications of all of the above in a <u>section on assumptions and traffic.</u>

SIEMENS

Pattern: Resource Pool Size Requirement

- Often omitted.
- Driven by average and peak transaction rates, resource holding times.
- Depends on technology, implementation, number of users to be supported.
- Depends on exhaustion probability we can tolerate, which in turn depends on the service requirements (connection with reliability!).
  - Ex: 1.0E-12<p<1.0E-06.

SIEMENS

Antipattern: Specified Resource Utilization

"The CPU utilization shall be 60%."

WRONG because:

- The resource utilization depends on the hardware as well as on the traffic.
  - The hardware could be changed
  - Light load => low utilization, so why 60%? What about peak load?

BUT: the measure is of interest:

- Resource utilization is important to the extent that it has a direct impact on service:
  - High utilization limits the system's ability to cope with short-term variations in traffic
  - High utilization at the desired peak load could lead to large response times
  - Very low utilization: system over-engineered or engineered to provide room for growth?

SIEMENS

What About Placing an Upper Limit on Average Resource Utilization?

- The average response time will rise as the average utilization of the bottleneck resource increases
- If the average utilization of the bottleneck resource is too high, performance will be adversely affected if there is a spike in traffic.

➢ Sometimes appropriate to state that "The average utilization of resource X shall be less than Y% in the peak hour."

- ➢ Y may be between 50 and 70%
- ➢ If Y refers to component that will carry double the load in case of failure, the average value of Y should be no more than 40% under normal operating conditions. *WHY?*

SIEMENS

Antipattern: "…all the time…"

- "The system shall be operational all the time."

Example:

- "I talk to General Petraeus all the time. I say 'all the time' -- weekly; that's all the time – …"

  -- President George W. Bush, June 28th, 2007.

http://www.prnewswire.com/cgi-bin/stories.pl?ACCT=104&STORY=/www/story/06-28-2007/0004617850&EDATE=

- *Problem: "…all the time…" is vague and not quantified.*

Do we mean

- *Continuously?*
- *99% of the time? 99.9% of the time?*
- *At regular intervals?*

SIEMENS

Antipattern: "…of the time"

"*The response time shall be less than 2 seconds 99% of the time."*

Problems:

- "…of the time" suggests time-averged statistics
- Response time is a *sample,* and so should be quantified with *sample statistics* (arithmetic mean etc.).

What does this mean?

- 99% of the transactions shall have a response time that is less than 2 secs?
- In any one-hour period, does it mean that response times observed during 0.99 x 3600 = 3564 sec shall be less than 2 seconds?

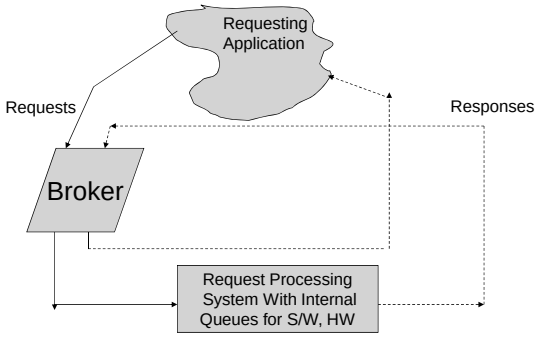
This is VAGUE!

SIEMENS

Exercise: An SOA-Based System

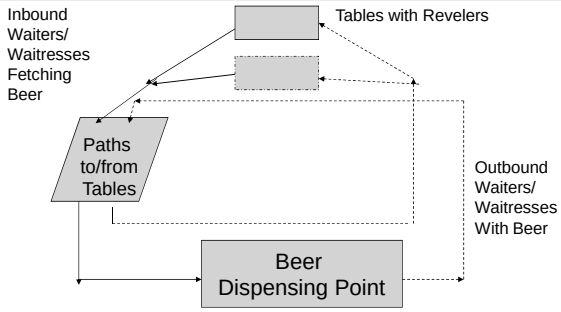
SIEMENS

Request/Response Interaction



SIEMENS

Oktoberfest



SIEMENS

Customer View

(Photos taken by the performance engineer shortly after opening time)

SIEMENS

Service View

SIEMENS

Exercise: Performance Requirements for Oktoberfest

- Different performance requirements for different stake holders:
  1. Customers/beer drinkers
  2. Brewery/tent owner
  3. Waiters and waitresses, represented by one or more union shop stewards
  4. Beer master, dispensing the beer
- One performance requirements team for each set of stake holders, representing its own interests independently

SIEMENS

Assignment: Define Performance Requirements

- Basic assumptions
- Customer view of performance
- System/service view of performance

Hints:

- Choose performance metrics carefully, while accounting for different view points.
  - Global view
  - Requirements on components; capacity of components
- Choose the variables carefully.

SIEMENS

Basic Assumptions, Physical Constraints

- Each server carries no more than 8 mugs of beer. [1 mug = 1 litre, weighs 1.25 kg empty]
- There are 8 customers per table
- Each server carries beer for only one table at a time.
- Each server is assigned to a constant set of tables, without help.
- Each server always goes to the same beer service point.
- Customers will drink an average 1.25 mugs of beer per hour, and never more than 2 mugs per hour

SIEMENS

Global/Customer View

- Identify the performance requirements as experienced by the customers.
- Identify the performance requirements as seen by the tent manager.
- Identify the performance requirements from the standpoint of the waiter/waitress.

Hints:

- Revenue (1 mug costs about €10)
- Look at the components
- Impatience

SIEMENS

System/Service View

- Identify performance requirements at the beer service point.
- Identify performance requirements on the pathways to/from the beer service point.
- First Service: 12:00
- Last Orders: 23:00

SIEMENS

Global View of Performance Requirements vs. Component View

Given:

- Performance requirements from a global standpoint

What performance requirements do the global requirements impose on components?

- Throughput and delay requirements on the beer service point
  - Physical constraint:
    - Rate at which beer pumped into a mug cannot exceed a certain level (Why not?)
  - Delay per waiter >= time to fill 8 mugs
- Delivery time bounded below by walking time from table to service point.
  - *Put a dedicated pathway in the architecture?*

SIEMENS

When the teams report back…

- Is each team's performance requirements consistent and compatible with those of other teams?
- How will conflicting requirements be resolved before the final requirements document is prepared and released?
  - Who is responsible for resolving those conflicts?
    - Architect?
    - Lead performance engineer?
  - What tradeoffs must be made when reconciling conflicting requirements?

SIEMENS

Reliability and Availability

SIEMENS

Reliability and Availability: General

Need to specify:

- Probability of failure (low) for system and components
- Fraction of time component, system is down
- Consequences of failures of specific types:
  - Specification of desired behaviour during failure scenarios
    - Failover or total shutdown?
    - Level of performance to be sustained during post-failure period? Response times? Permissible transaction loss rate?
    - How long may a failover take?
  - Automatic recovery from crash caused by overload if load abates within a set amount of time
- If using replication
  - How long should it take to replicate a specific piece of data after it is updated?
  - Is data replication ongoing or batched? Performance requirements?

SIEMENS

Testing for Reliability is Difficult

- Looking for the occurrence of events (failures) that should not happen often.

Two approaches:

- Soak test: run for a long time at varying loads and see what happens.
  - *Hard to define "long enough."*
- Look for warning signs in measurement data from performance tests
  - Memory leaks, erratic resource consumption under constant demand, …

SIEMENS

Look for Warning Signs in Performance Data

- Identify characteristics of system **stability** *under constant loads* and **require that they occur.**
  - Resource usage (CPU busy, bandwidth, memory, number of open TCP sockets, etc.) should be fairly constant under constant load.
  - Average response time should vary little under constant load.
- Identify characteristics of system measurements known to indicate poor reliability or later failure, and **require that they not occur**.
  - Increasing occupancy of any discrete resource (pools) => ***leak***
  - *Response time and/or utilization increasing with time*

SIEMENS

Managing Performance Requirements

SIEMENS

Managing the Performance Requirements

- Requires attention throughout the life cycle.
- Designated owner of the performance requirements is essential:
  - Tracks performance requirements, handles
    - Change control
    - Traceability
    - Links to business and engineering needs
  - Mediates between stakeholders when performance requirements are being negotiated and written
    - Architects, designers, testers, marketers, sales engineers, even lawyers (!)
- Performance requirements should be centrally stored to facilitate viewing by all stakeholders

SIEMENS

Architecture and Performance Requirements

- The performance requirements should affect the choice of system architecture.
- Once the architecture has been chosen, the overall performance requirements will affect the performance requirements of the individual system components.
- The performance requirements must be tracked throughout the software life cycle.
  - ❖ Changes must be reflected in the architecture and in the design.

SIEMENS

Performance Requirements and Performance Testing

- For verification, must be able to obtain the metrics stated in the requirement through measurement.
  - Performance tester and requirements writers should communicate while the requirements are being written
- Structure performance tests to allow verification of multiple performance requirements simultaneously.
- Example:
  - Measure average and maximum response times
  - Measure resource utilizations to ensure that the load is sustainable (also keeps response time from being too variable).
  - Check for constant averages under constant load
    - Averages variable or exhibiting a trend => problem!

SIEMENS

References

- IEEE Recommended Practice for Software Requirements Specifications, IEEE Std 830-1998
- Lilja, D. J. *Measuring Computer Performance: a practitioner's guide.* Cambridge University Press, 2000.
- Nixon, B.A.: Managing performance requirements for information systems. *Proc. First WOSP 1998,* pp.131-144, 1998.
- R. F. Rey (ed)., *Engineering and Operations in the Bell System*, AT&T Bell Laboratories, 1983.
- Smith, Connie U., and Lloyd G. Williams. *Performance Solutions: a Practical Guide to Creating Responsive, Scalable Software.* Addison-Wesley, 2000.