中国科学技术大学
University of Science and Technology of China

Software Architecture

SSE USTC　Qing Ding
dingqing@ustc.edu.cn
http://staff.ustc.edu.cn/~dingqing

Quality Attributes of Architecture II
Performance

Agenda

- Performance
  - Source of stimulus
  - Stimulus
  - Environment
  - Artifact
  - Response
  - Response measure
  - Tactics

PERFORMANCE

- Performance is about timing
  - Basically performance is concerned with how long it takes the system to respond when an event occurs.
  - One of the things that make performance complicated is the number of event sources and arrival patterns.
    - Events can arrive from user requests, from other systems, or from within the system.
    - An arrival pattern for events may be characterized as either periodic or stochastic. Events can also arrive sporadically.
    - Multiple users or other loading factors can be modeled by varying the arrival pattern for events.

PERFORMANCE

- Performance is about timing
  - The response of the system to a stimulus can be characterized by
    - latency (the time between the arrival of the stimulus and the system's response to it)
    - deadlines in processing (in the engine controller, for example, the fuel should ignite when the cylinder is in a particular position, thus introducing a processing deadline)
    - the throughput of the system (e.g., the number of transactions the system can process in a second)
    - the jitter of the response (the variation in latency)
    - the number of events not processed because the system was too busy to respond
    - the data that was lost because the system was too busy

PERFORMANCE

- Source of stimulus.
  - The stimuli arrive either from external (possibly multiple) or internal sources.
- Stimulus.
  - The stimuli are the event arrivals. The arrival pattern can be characterized as periodic, stochastic, or sporadic.
- Artifact.
  - The artifact is always the system’s services.
- Environment.
  - The system can be in various operational modes, such as normal, emergency, or overload.

PERFORMANCE

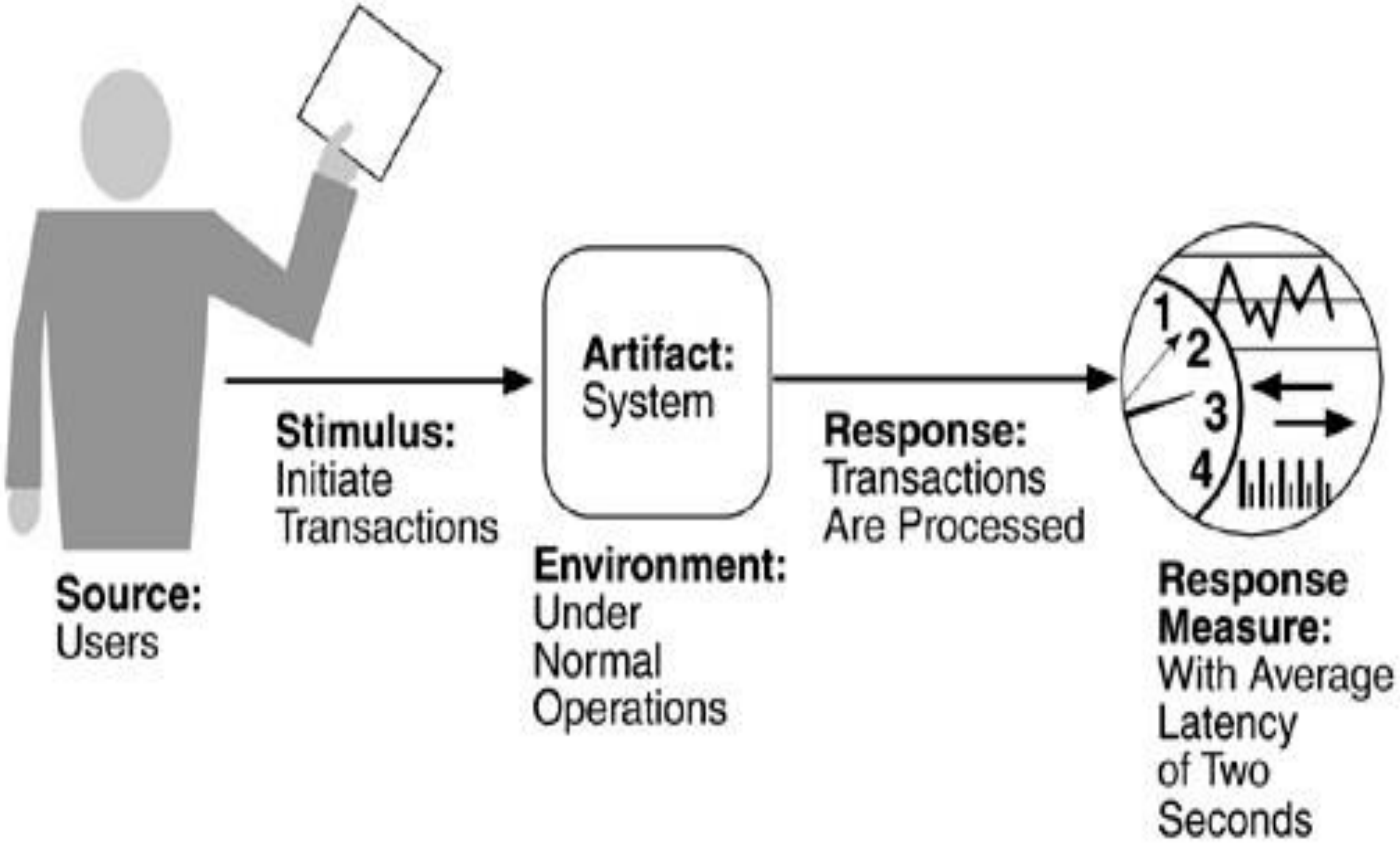
- Response.
  - The system must process the arriving events. This may cause a change in the system environment (e.g., from normal to overload mode).

- Response measure.
  - The response measures are the time it takes to process the arriving events (latency or a deadline by which the event must be processed), the variation in this time (jitter), the number of events that can be processed within a particular time interval (throughput), or a characterization of the events that cannot be processed (miss rate, data loss).

PERFORMANCE

Portion of Scenario	Possible Values
Source	One of a number of independent sources, possibly from within system
Stimulus	Periodic events arrive; sporadic events arrive; stochastic events arrive
Artifact	System
Environment	Normal mode; overload mode
Response	Processes stimuli; changes level of service
Response Measure	Latency, deadline, throughput, jitter, miss rate, data loss

PERFORMANCE



Source: Users

Stimulus: Initiate Transactions

Artifact: System

Environment: Under Normal Operations

Response: Transactions Are Processed

Response Measure: With Average Latency of Two Seconds

跳过

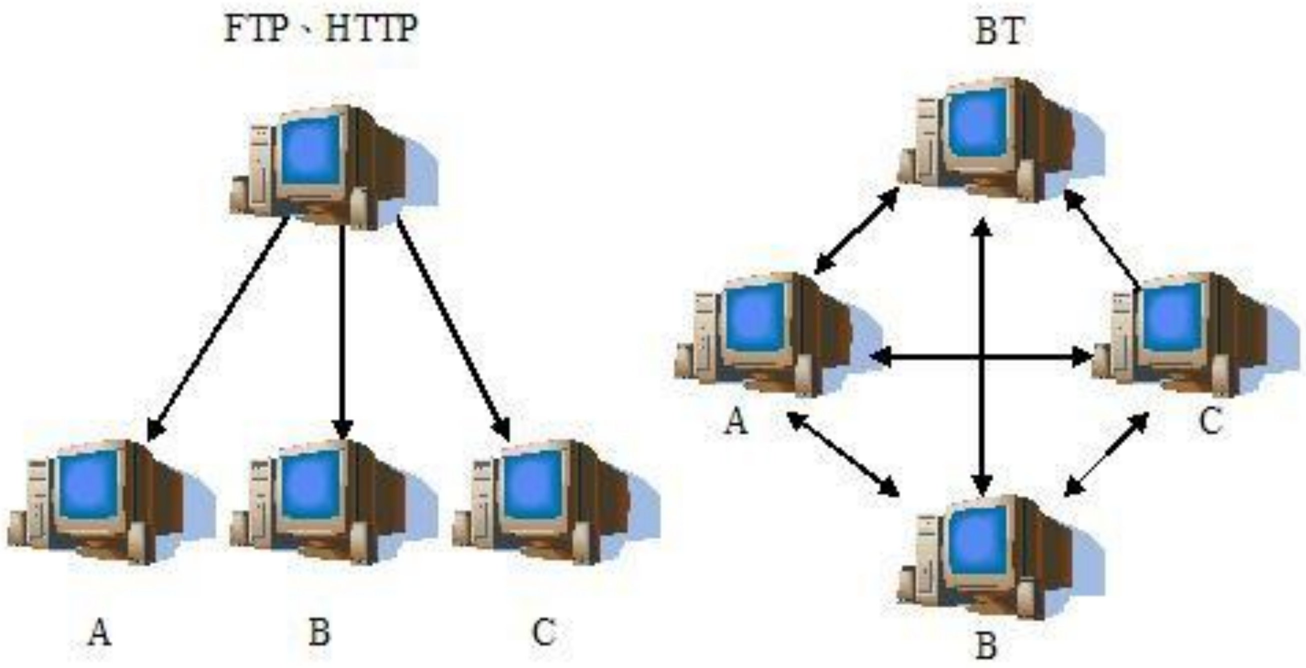
BT – A case of performance optimization

- BitTorrent, in short BT
  - Developed by Bram Cohen in 2002
  - Bram Cohen entered SNYU in 1993

- Bram released BT in 2002, but it was not so attractive at that time
- In May, 2003, Bram published a 5---page paper on Workshop on Economics of Peer---to---Peer Systems, 2003
  - This paper has been referenced more than several hundred times
  - Now, BT and its various modified versions, such as BitComet and BitSpirit have been an important common tool for users.

BT

- High speed download



BT

- FTP
  - The seed can not be offline until all the clients finish downloading the file
  - The cost of bandwidth is unilateral. The upload bandwidth of seed will be the bottleneck
- BT
  - Once a client accomplishes download the whole seed, it is a new seed. The original seed could be offline then.
  - The original seed even can be offline when no new seed generated.
  - If the seed divides the file into 10 segments, and sends segments 1---3 to A, 4---6 to B, and 7--- 10 to C, then the seed can switch from online to offline, since A, B and C can download segments from each other.
- BT reduces the workload of seeds and prolongs the life of seeds
  - If the original seed leaved, and one of A, B, C voluntarily to be a new seed, the file still be downloadable.

BT

- The components of BT
  - Tracker
    - Records the name and network location of all the downloading members
  - Torrent
    - Records the location of Tracker and the names of all the segments
    - A new member needs to obtain a torrent at first and then get the tracker by the information retrieved from torrent.
  - Rarest First Policy
    - Every member will share the rarest segments to others in order to speed up the download.
    - For example, if user A has segments 1, 2, 3, ,user B has segments 2, 4, 5, user C has segments 1, 3, 5, then the rarest segment is 4. If a member wants to share segments with user B, B will prefer to send segment 4 to him.

BT

- The components of BT
  - Choking Policy
    - If a file is very popular, there will be thousands of users want to download it.
    - User A is one of the members who are downloading this file. He must receive many request to share segments.
    - But his bandwidth is limited. He only can concurrent share segments to 4 other users.
    - He has to refuse most of the request. How to refuse?
    - Choking policy: unchoke the 4 fastest users who are sharing segments to me.
  - Optimistic Unchoking
    - For every 30 seconds, BT randomly selects a user and sends data to him.
    - To discover the potential fast member.

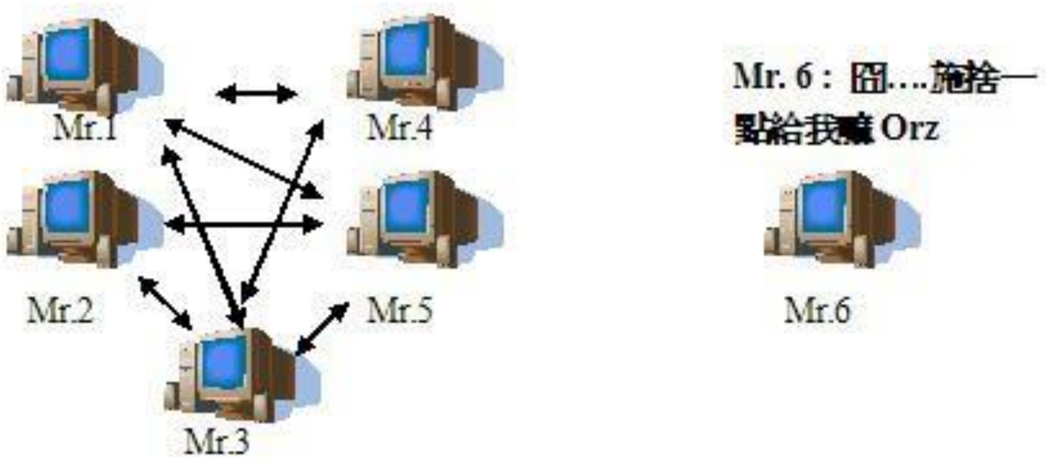
BT – 性能优化实例

- Dissecting BitTorrent：Five Months in a Torrent’s Life
  - （1） BT檔案往往會有flash crowd情形：分享開始的前幾天會湧入大 量人潮，然而高潮退去後人也散得快。
  - （2） 下載的速度呈現「多數人下載慢，極少數人下載速度超快」的 情形；不過即使「多數人下載慢」，下載的平均速度仍然比ftp快上不 少。根據實驗，平均下載速度是240K，90%的人速度不會超過520K。 有極少數的人下載速度會達到每秒3000K以上。
  - （3） 檔案的存活天數難以從檔案分享十天後的狀況來預測：紐西蘭 學生嘗試著去預測檔案存活天數，不過失敗了。後面我們會看到另外 一篇paper是如何成功預測的。

BT

- Modeling and Performance Analysis of BitTorrent---Like Peer---to---Peer Networks
  - Dongyu Qiu and R. Sirkant, UIUC
  - In 2004, SIGCOMM

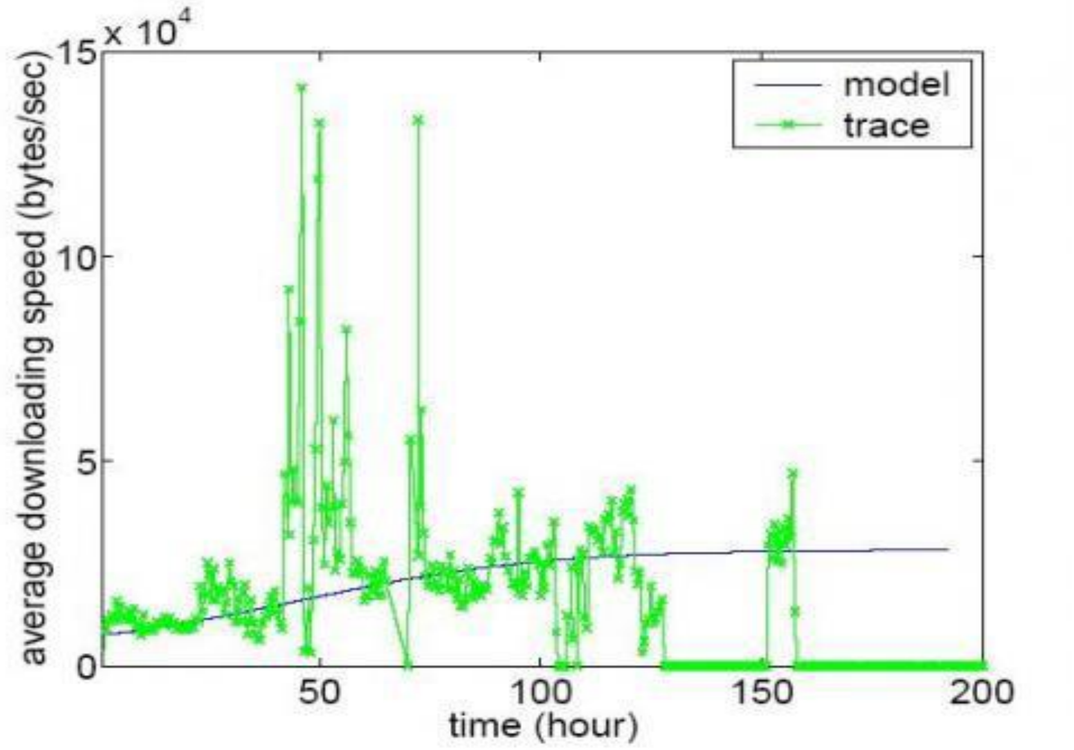
  - Fluid Model
  - Free Riding



（註：本圖絕對未影射任何特定人、事、時、地、物...請各位讀者不要亂想 XD）

BT

- Measurements, Analysis, and Modeling of BitTorrent---like Systems.
  - The fastest speed of downloading: 50 hours after releasing of file

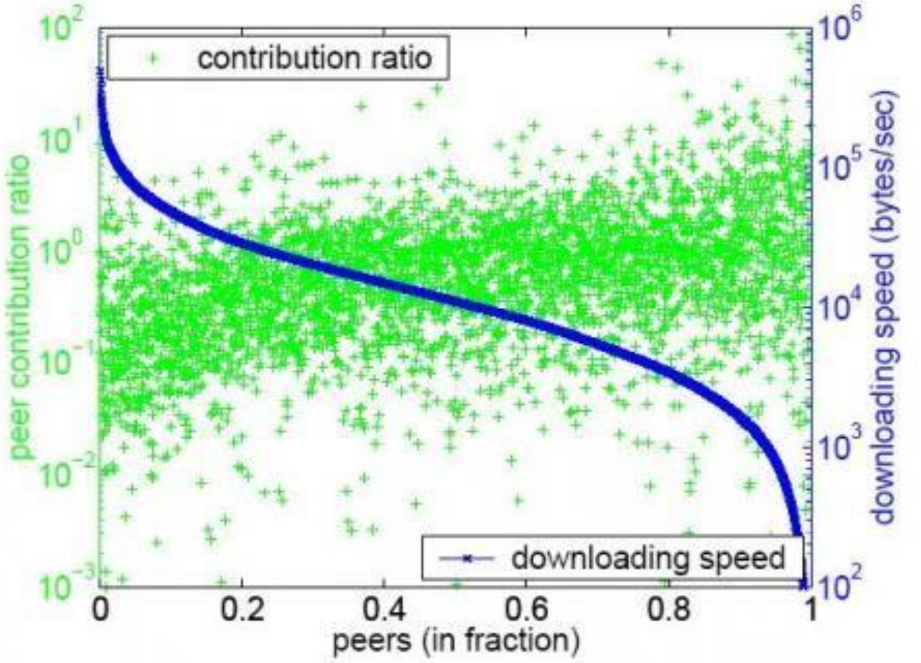


(b) Downloading speed

BT

- Measurements, Analysis, and Modeling of BitTorrent---like Systems.

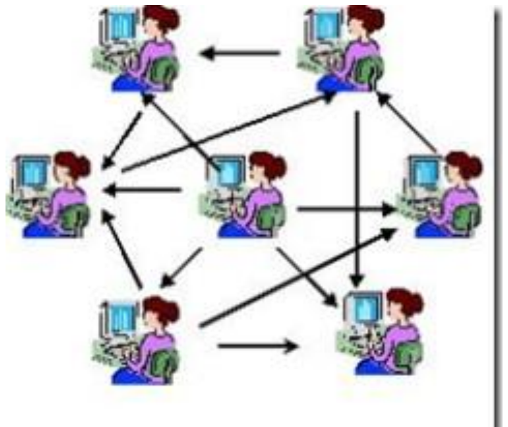
– A flaw of BT: speed of downloading doesn’t reflect the difference among speed of uploading



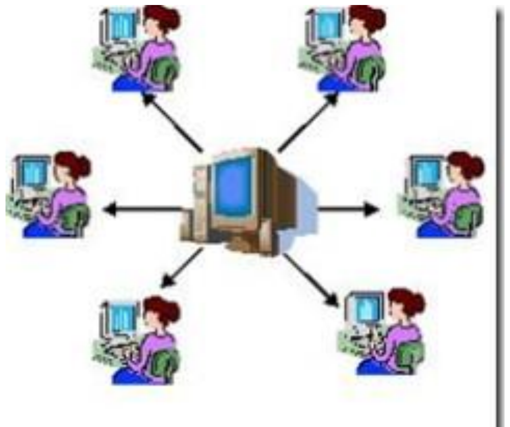
(a) The peer downloading speed and contribution ratio

BT

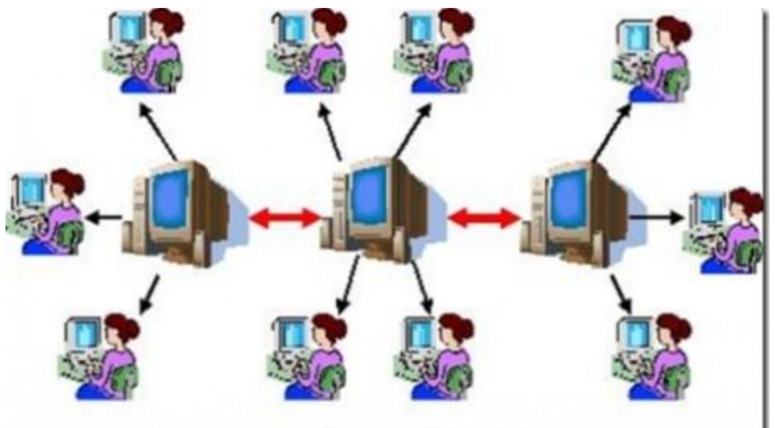
- Why don’t BT provide searching?



Anthill 的搜尋方式



Napster 的搜尋方式



KaZaA, eMule 的搜尋方式

Performance Analysis

- An Analytical Model for Multi---Tier Internet Services and its Applications
  - Bhuvan Urgaonkar, Giovanni Pacificiy, Prashant Shenoy, Mike Spreitzery, and Asser Tantawiy

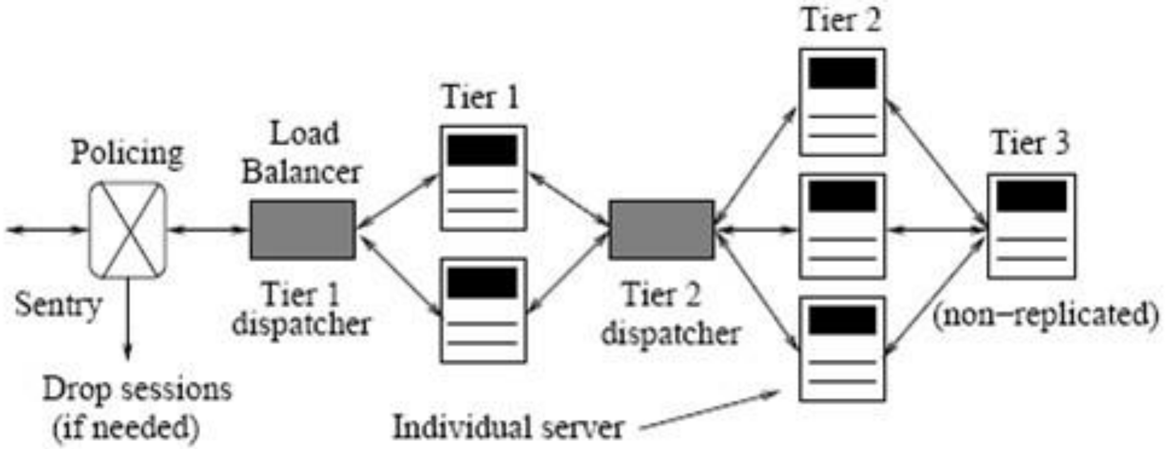
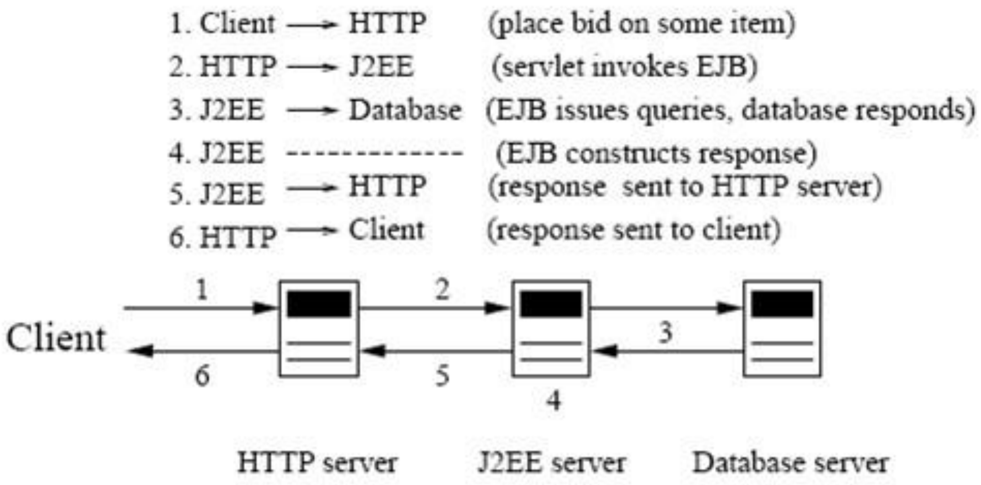


Figure 1: A three-tier application.

Performance Analysis

- An Analytical Model for Multi---Tier Internet Services and its Applications

– Bhuvan Urgaonkar, Giovanni Pacificiy, Prashant Shenoy, Mike Spreitzery, and Asser Tantawiy

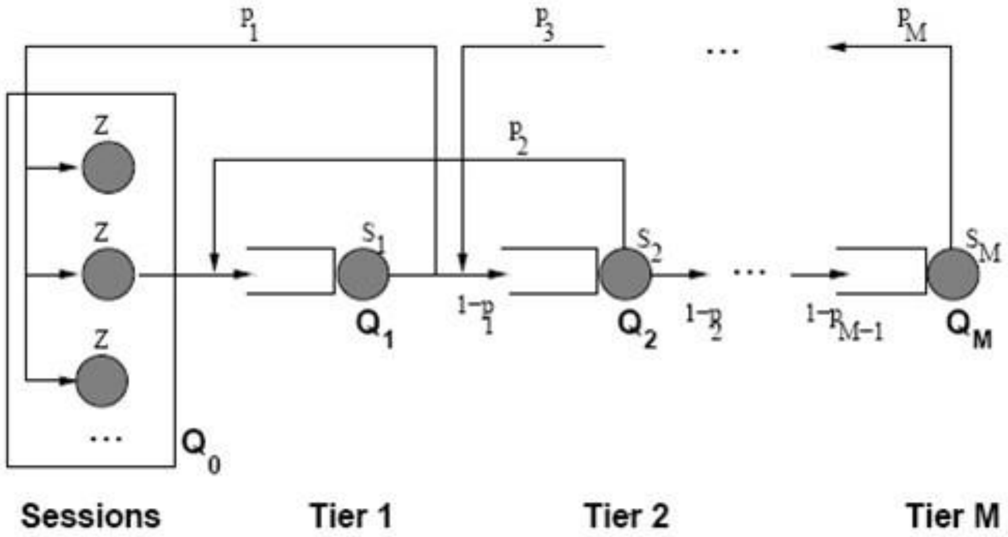


1. Client → HTTP (place bid on some item)
2. HTTP → J2EE (servlet invokes EJB)
3. J2EE → Database (EJB issues queries, database responds)
4. J2EE ------------- (EJB constructs response)
5. J2EE → HTTP (response sent to HTTP server)
6. HTTP → Client (response sent to client)

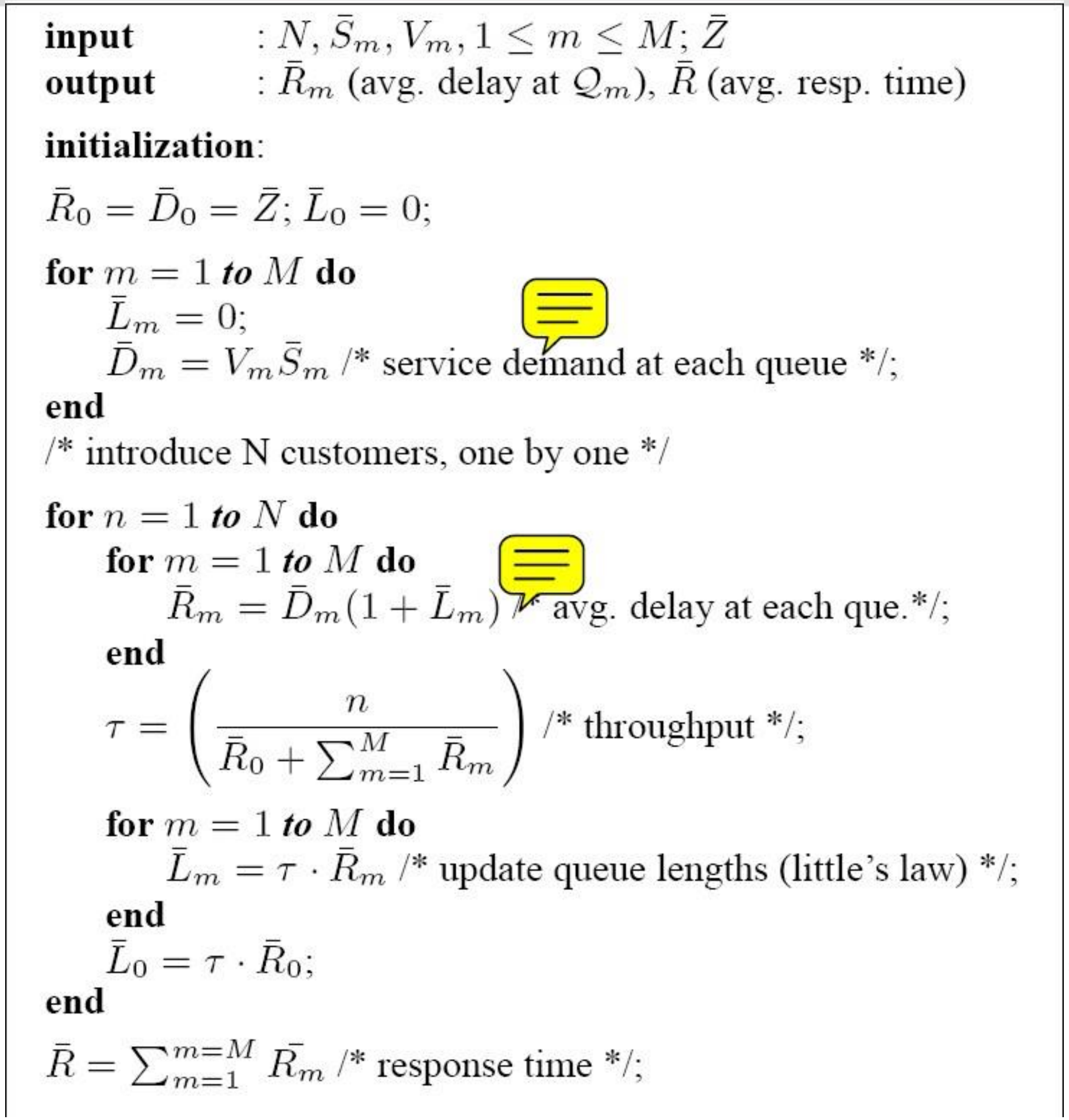
Figure 2: Request processing in an online auction application.

Performance Analysis

- An Analytical Model for Multi---Tier Internet Services and its Applications
  - Bhuvan Urgaonkar, Giovanni Pacificiy, Prashant Shenoy, Mike Spreitzery, and Asser Tantawiy

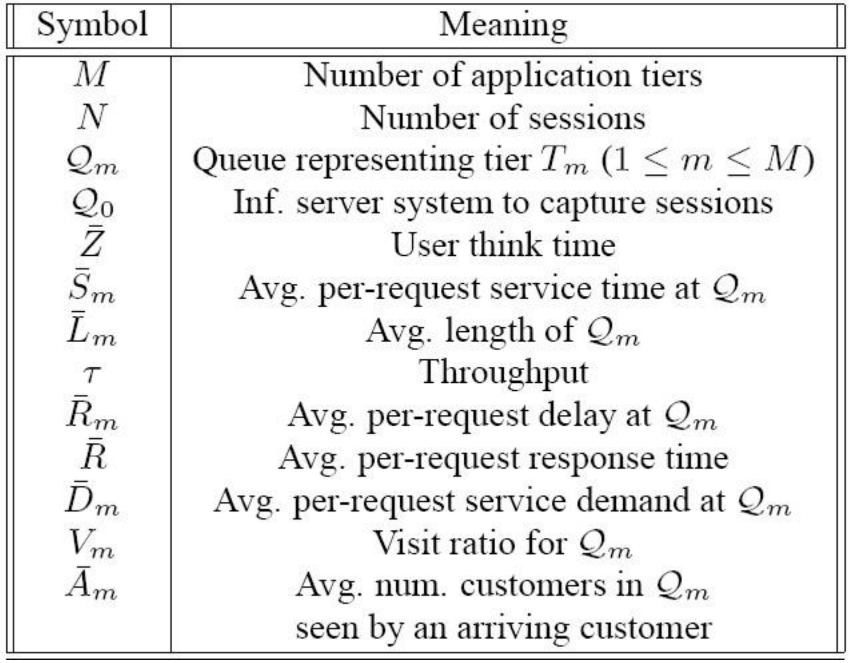


Performance Analysis



```
input          : $N, \bar{S}_m, V_m, 1 \leq m \leq M; \bar{Z}$
output         : $\bar{R}_m$ (avg. delay at $\mathcal{Q}_m$), $\bar{R}$ (avg. resp. time)

initialization:
$\bar{R}_0 = \bar{D}_0 = \bar{Z}; \bar{L}_0 = 0;$
for $m = 1$ to $M$ do
    $\bar{L}_m = 0;$
    $\bar{D}_m = V_m \bar{S}_m$ /* service demand at each queue */;
end
/* introduce N customers, one by one */
for $n = 1$ to $N$ do
    for $m = 1$ to $M$ do
        $\bar{R}_m = \bar{D}_m (1 + \bar{L}_m)$ /* avg. delay at each que.*/;
    end
    $\tau = \left( \frac{n}{\bar{R}_0 + \sum_{m=1}^{M} \bar{R}_m} \right)$ /* throughput */;
    for $m = 1$ to $M$ do
        $\bar{L}_m = \tau \cdot \bar{R}_m$ /* update queue lengths (little's law) */;
    end
    $\bar{L}_0 = \tau \cdot \bar{R}_0;$
end
$\bar{R} = \sum_{m=1}^{m=M} \bar{R}_m$ /* response time */;
```

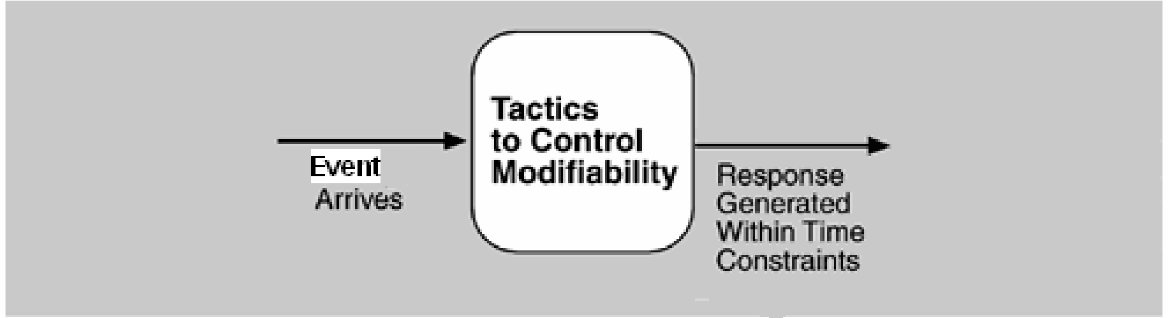


Symbol	Meaning
M	Number of application tiers
N	Number of sessions
$\mathcal{Q}_m$	Queue representing tier T_m $(1 \leq m \leq M)$
$\mathcal{Q}_0$	Inf. server system to capture sessions
$\bar{Z}$	User think time
$\bar{S}_m$	Avg. per-request service time at $\mathcal{Q}_m$
$\bar{L}_m$	Avg. length of $\mathcal{Q}_m$
τ	Throughput
$\bar{R}_m$	Avg. per-request delay at $\mathcal{Q}_m$
$\bar{R}$	Avg. per-request response time
$\bar{D}_m$	Avg. per-request service demand at $\mathcal{Q}_m$
V_m	Visit ratio for $\mathcal{Q}_m$
$\bar{A}_m$	Avg. num. customers in $\mathcal{Q}_m$ seen by an arriving customer

Performance Tactics

- The goal of performance tactics is to generate a response to an event arriving at the system within some time constraint.
- The event can be single or a stream and is the trigger for a request to perform computation.
- It can be the arrival of a message, the expiration of a time interval, the detection of a significant change of state in the system's environment, and so forth.
- The system processes the events and generates a response. Performance tactics control the time within which a response is generated.
- Latency is the time between the arrival of an event and the generation of a response to it.

Goal of performance tactics


Event Arrives
Tactics to Control Modifiability
Response Generated Within Time Constraints

Performance Tactics

- After an event arrives, either the system is processing on that event or the processing is blocked for some reason.

- This leads to the two basic contributors to the response time: resource consumption. and blocked time

- Resource consumption.
  - Resources include CPU, data stores, network communication bandwidth, and memory, but it can also include entities defined by the particular system under design.
  - Events can be of varying types (as just enumerated), and each type goes through a processing sequence.

Performance Tactics

- Blocked time.
  - A computation can be blocked from using a resource because of contention for it, because the resource is unavailable, or because the computation depends on the result of other computations that are not yet available.
    - Contention for resources.
    - Availability of resources.
    - Dependency on other computation.

- With this background, we turn to our three tactic categories: resource demand, resource management, and resource arbitration.

Performance Tactics-resource demand

- Event streams are the source of resource demand.
  - Two characteristics of demand are the time between events in a resource stream (how often a request is made in a stream) and how much of a resource is consumed by each request.

- One tactic for reducing latency is to reduce the resources required for processing an event stream. Ways to do this include the following.
  - Increase computational efficiency.
  - Reduce computational overhead.

- Another tactic for reducing latency is to reduce the number of events processed. This can be done in one of two fashions.
  - Manage event rate
  - Control frequency of sampling.

无线传感器例子：减少事件到达率，降低采用率

Performance Tactics-resource demand

- Other tactics for reducing or managing demand involve controlling the use of resources.
  - Bound execution times. Place a limit on how much execution time is used to respond to an event. Sometimes this makes sense and sometimes it does not. For iterative, data---dependent algorithms, limiting the number of iterations is a method for bounding execution times.
  - Bound queue sizes. This controls the maximum number of queued arrivals and consequently the resources used to process the arrivals.

Performance Tactics-resource management

- Even though the demand for resources may not be controllable, the management of these resources affects response times.

- Some resource management tactics are:
  - Introduce concurrency.
  - Maintain multiple copies of either data or computations.
  - Increase available resources.

Performance Tactics-resource arbitration

- Whenever there is contention for a resource, the resource must be scheduled. Processors are scheduled, buffers are scheduled, and networks are scheduled.

- The architect's goal is to understand the characteristics of each resource's use and choose the scheduling strategy that is compatible with it.

- A scheduling policy conceptually has two parts: a priority assignment and dispatching.
  - Competing criteria for scheduling include optimal resource usage, request importance, minimizing the number of resources used, minimizing latency, maximizing throughput, preventing starvation to ensure fairness, and so forth. The architect needs to be aware of these possibly conflicting criteria and the effect that the chosen tactic has on meeting them.

调度：1、mapping 空间上（把任务、资源分配到哪个机器）
　　　2、ordoring时间上

Performance Tactics-resource arbitration

- A high---priority event stream can be dispatched only if the resource to which it is being assigned is available. Sometimes this depends on pre----empting the current user of the resource.
  - Possible preemption options are as follows: can occur anytime; can occur only at specific pre---emption points; and executing processes cannot be pre---empted.
- Some common scheduling policies are:
- First---in/First---out.
  - FIFO queues treat all requests for resources as equals and satisfy them in turn.
  - One possibility with a FIFO queue is that one request will be stuck behind another one that takes a long time to generate a response.
  - As long as all of the requests are truly equal, this is not a problem, but if some requests are of higher priority than others, it is problematic.

Performance Tactics-resource arbitration

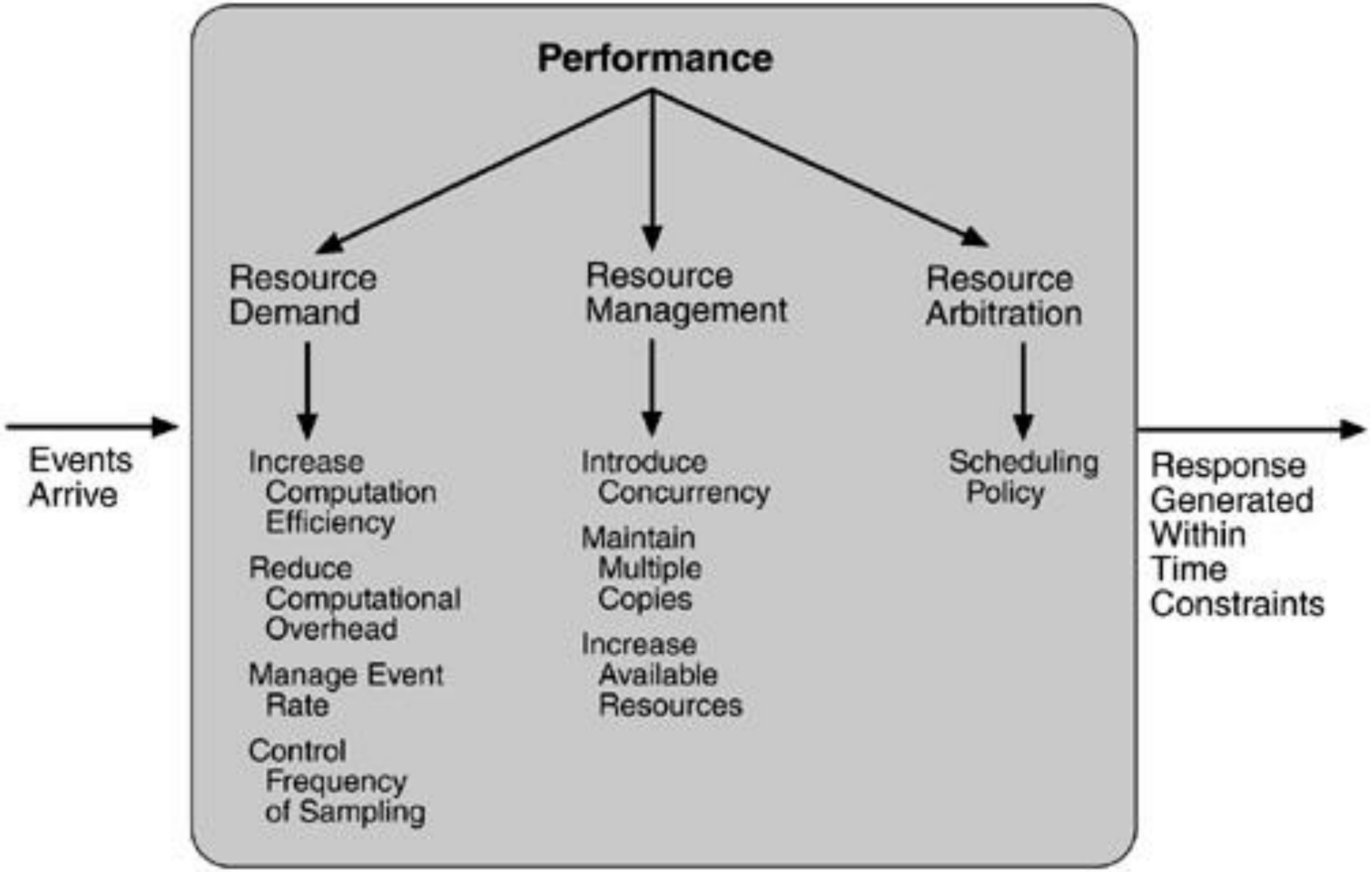
- Fixed---priority scheduling.
  - Fixed---priority scheduling assigns each source of resource requests a particular priority and assigns the resources in that priority order.
  - This strategy insures better service for higher---priority requests but admits the possibility of a low---priority, but important, request taking an arbitrarily long time to be serviced because it is stuck behind a series of higher---priority requests.
  - Three common prioritization strategies are
    - semantic importance
    - deadline monotonic.
    - rate monotonic.

Performance Tactics-resource arbitration

- Dynamic priority scheduling:
  - round robin. Round robin is a scheduling strategy that orders the requests and then, at every assignment possibility, assigns the resource to the next request in that order. A special form of round robin is a cyclic executive where assignment possibilities are at fixed time intervals.
  - earliest deadline first. Earliest deadline first assigns priorities based on the pending requests with the earliest deadline.

- Static scheduling. A cyclic executive schedule is a scheduling strategy where the pre---emption points and the sequence of assignment to the resource are determined offline.

Performance Tactics-Summary


Performance
Resource Demand
Resource Management
Resource Arbitration
Events Arrive
Increase Computation Efficiency
Reduce Computational Overhead
Manage Event Rate
Control Frequency of Sampling
Introduce Concurrency
Maintain Multiple Copies
Increase Available Resources
Scheduling Policy
Response Generated Within Time Constraints

Project

- Suppose you needs to improve the performance of your SNS website. Please describe your design from the following aspects:
  - If your SNS website is deployed into a cluster, what tactics would you want to adopt to improve utilization of computing resource?
  - What tactics do you think is suitable for your SNS website to reduce the resource contention?
  - According to the functions your designed for your SNS website, please give your scheduling policy for resource arbitration.