Lecture 2

Performance Evaluation Process, Models and Metrics

- Usage
- Function
- Model
- Metrics
- Examples

Capacity Planning Usage

- Current system, new system
- HW
- SW
  - OS
  - Applications
- Measurement of existing system
- Tuning current system
- Planning for future systems
- See Figs on bad planning

Capacity Planning Basic Methods

- Measurement
- Modeling
  - Solution methods for models
    - analytical, simulation, mixed
    - operational analysis, approximations
  - Parameter estimation
    - existing systems, future systems
    - guesswork
    - workload modeling

Capacity Planning Example Usages

- Why is my machine so slow?
  - would 64MB extra memory help?
  - should I put the 64MB in main memory or into the display card?
  - what if I just change the scheduling algorithm?
- Is Pentium II fast enough for this server, or do we need to use a Pentium IV?
  - how fast Pentium IV?
  - what about 2 years from now?

Capacity Planning Example Usages

- What about the new system?
  - Is it fast enough? What does ”fast” mean?
  - Is it balanced?
    - slow component => everybody is slowed down
    - fast component => waste of money
- What about the current system?
  - How do we get most of it out with the least expenses?
  - Can we modify it or do we need completely new system? When do we need it?

Example: Bank Application

[Menasce 94]

- System: terminals, network, CPU, 2 disks
- Service
  - Queries, 70% of transactions, max resp. time 3 s
  - Updates into many files, max resp. time 10 s.

job classes

require-ment

- measured service time per transaction

	Quer	Upd
CPU	0.20	0.30 sec
Disk1	0.30	0.80
Disk2	0.25	0.45

measu-rement

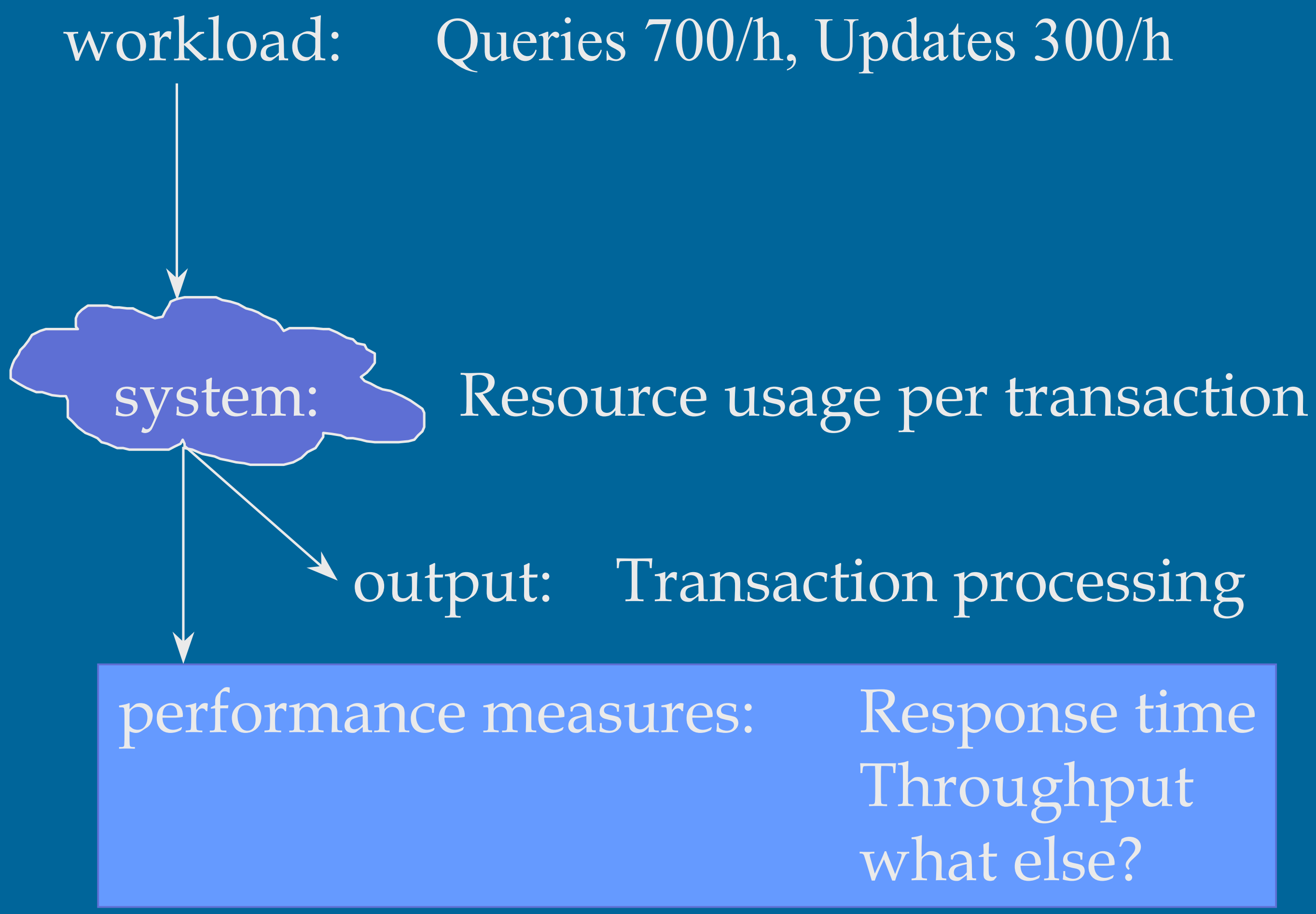
- Query resp. time 2.3 s, Update resp. time 9 s
- Queries 700/h, Updates 300/h

work load

- Can the system handle it, if the query rate goes up 30%?

future work load

Example (contd)



workload: Queries 700/h, Updates 300/h

system: Resource usage per transaction

output: Transaction processing

performance measures: Response time
Throughput
what else?

Saturation

- System is saturated, if the performance requirement for some job class is not met
  - e.g., response time > 3 s
  - *no* device is necessarily saturated
- A device is saturated if a physical device is at use close to 100% of the time
  - CPU utilization is close to 100%?
  - network is close to 100% utilized
  - response times very high, system is saturated
  - *many* devices may be saturated

Performance Metrics

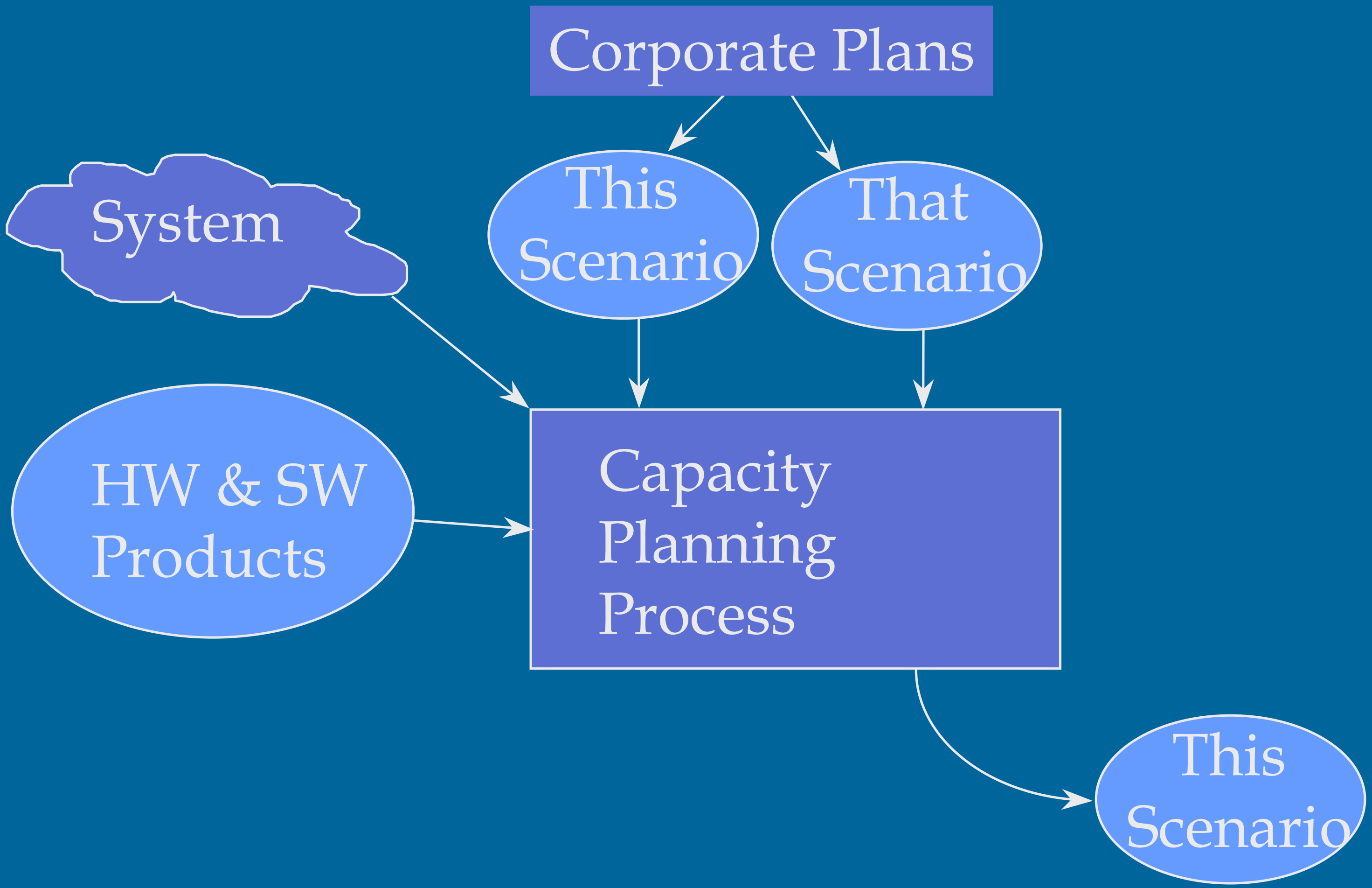
- Customer View, External Performance
  - response time, turnaround time, reaction time
  - throughput, flow
  - availability

Bottom line?
Goal?

- System View, Internal Performance
  - response time (R, Ri)
  - throughput (X, Xi)
  - utilization (U, Ui)
  - queue length (Q, Qi)
  - system capacity?
  - component capacity?
  - cost

for system
for each device i

Function of Capacity Planning Process



Corporate Plans

This Scenario

That Scenario

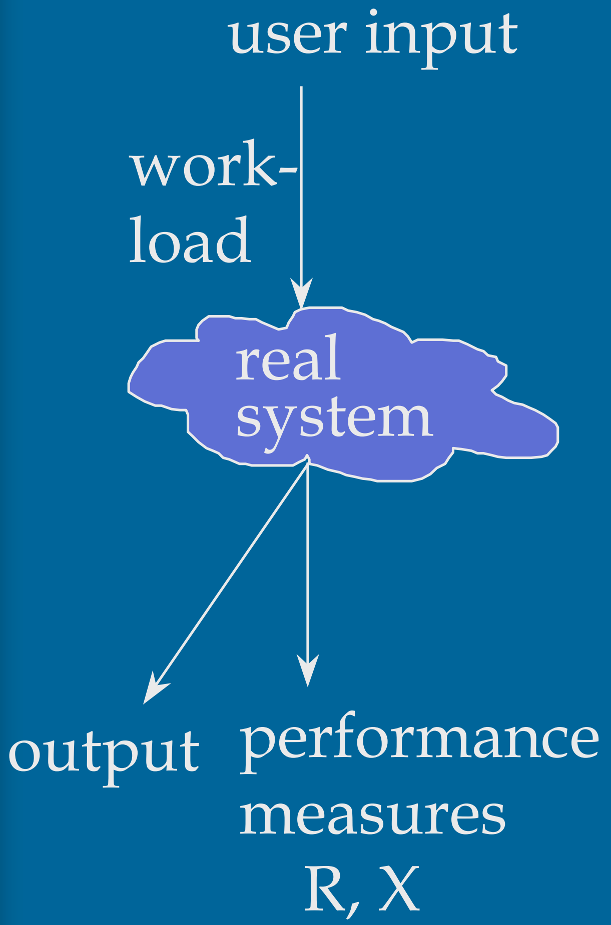
System

HW & SW Products

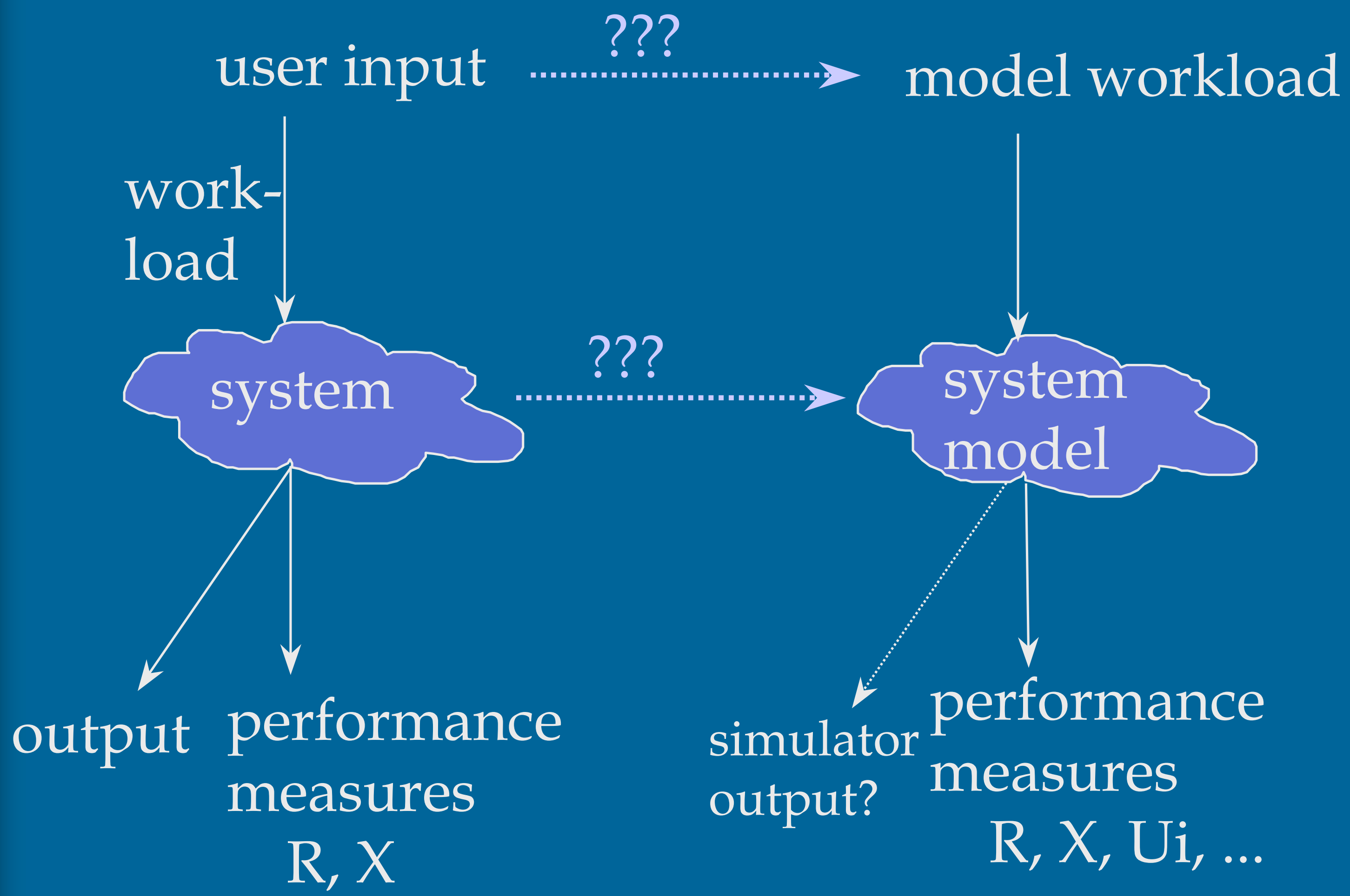
Capacity Planning Process

This Scenario

System Model (2)



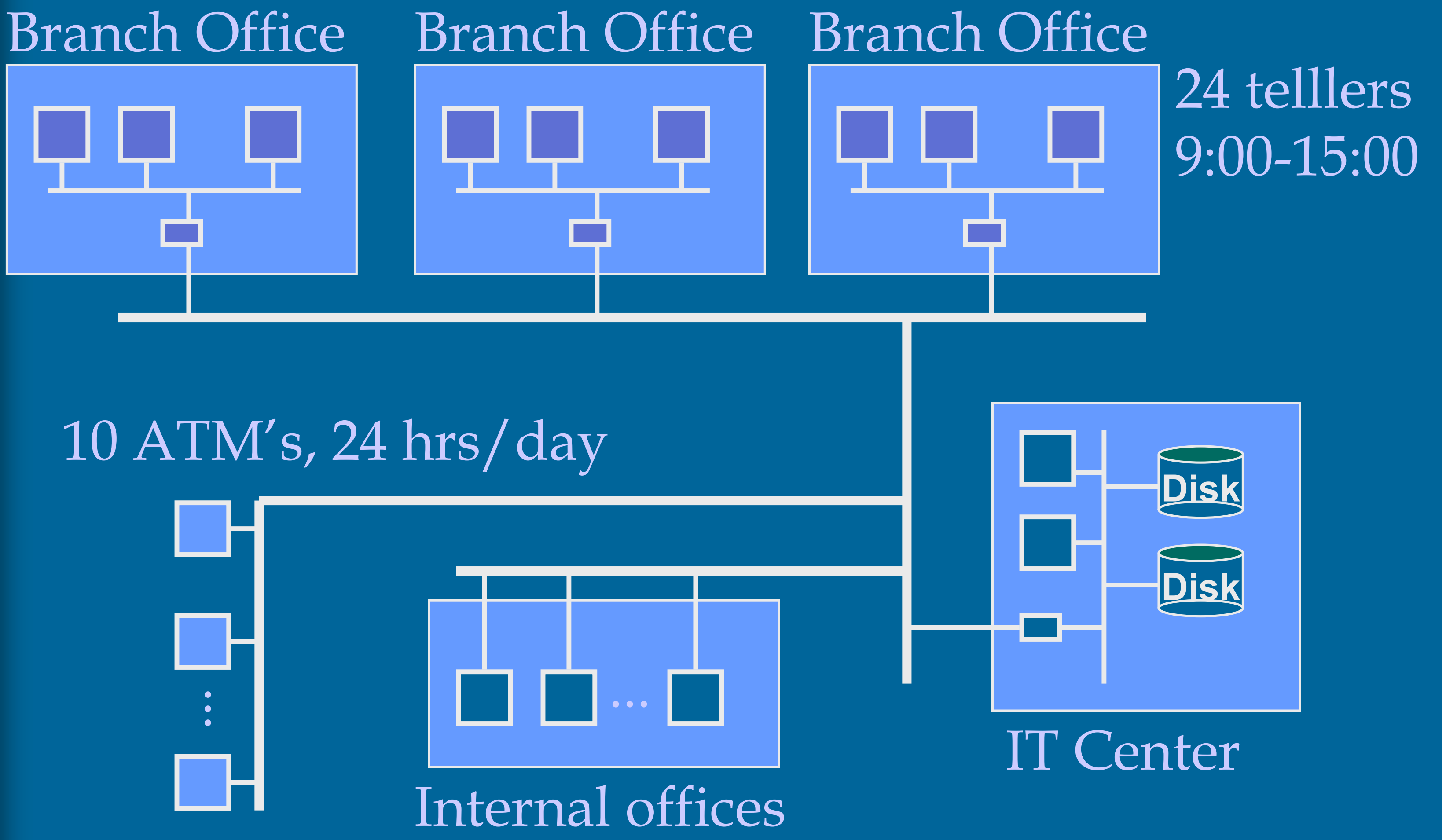
System Model (2)



Example on Prediction

- Previous CPU utilization
  - Table 1.2 [Menasce 94]
- Linear forecast of CPU utilization
  - Table 1.3 [Menasce 94]
- Bad estimate for September. Why?
  - bad assumption: linear growth
  - possible changes in workload not considered
  - CPU utilization might be bad metric for system performance
    - Better: response time? for different job classes?

Example Problem: Bank



Teller Load to System

- 2 online transactions per customer
- peak 11:30-13:30: 20 customers/hour

  I.e., 24 * 20 * 2 = **960 trans/h** (total), or

  320 trans/h (per branch) or

  80 trans/h (per teller), or
- other: 12 customers/h

  I.e., 24 * 12 * 2 = **576 transactions/h** (total)

ATM Load to System

- 1.2 transactions/customer **(in average)**
- peak 8:00-9:00, 15:00-21:00
  15 customers/h, I.e.,
  10*15*1.2 = **180 trans/h** (total)
  or **18 trans/h** (per ATM)
- other: 7.5 cust/h, I.e., **90 trans/h** (total)

Average System Response Time

• Teller	peak 1.23 s	limit 3 s.
• ATM	peak 1.02 s	limit 4 s.

Expansion?

- Teller peak load is 960 trans/hr
  New branch office per every 2 months:
  320 new trans/h per 2 months, I.e.,
  160 new trans/h per month, I.e.,
  teller peak estimate: **960 + 160m trans/h**

months

- ATM peak load is 180 trans/h
  20 new ATMs per 2 months, I.e.,
  10 * 18 = 180 new trans/hr/month, I.e.,
  ATM peak estimate: **180+180m trans/h**

Expansion Questions

- How long are resp. times OK?
  R(teller) < 3 sec? R(ATM) < 4 sec?
- What upgrade is needed and when?
  - new CPU? new disks? new traffic controller?
  - Figs 1.4 and 1.3 [Menasce 94]
- Would another, distributed approach be better?
  - more scalable?
  - Figs 1.5 and 1.6 [Menasce 94]

Performance Metrics, Customer View, External Performance

• Response time	(vasteaika)
• Turnaround time	(vastausaika)
• Reaction time	(reaktioaika)
• Throughput	(läpimenotiheys, -vuo)
• Availability	(käytettävyys)

Performance Metrics, System View, Internal Performance

Utilization (*) U	(käyttösuhde)
Queue length (*) Q	(jonon pituus)
Response time	(vasteaika)
Throughput	(läpimenotiheys)
Capacity(*)	(kapasiteetti)
Cost (*)	(hinta)

(*) per system, or per component