

About architecture design process

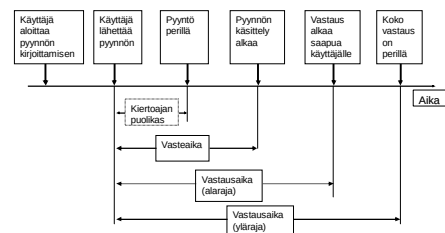
1. Basic functionality and composition
 - Roles, creators of data, processing needs, data flows, decisionmaking points
 - Modularity and evolvability, understandability
2. Addition of correct distribution mechanisms
 - Communication model
 - Decision model
 - Vertical and horizontal distribution
3. Performance analysis
 - Modeling, analysis → indications of the performance
4. Selection between alternative architectures

Suorituskyky: mittasuureet

- Palvelukyky (ability to serve)
 - vastausaika (response time)
 - kiertoaika, kääntymisaika (turnaround time)
 - vasteaika (reaction time)
 - suoritusteho, läpäisy (throughput)
- Käyttöaste (utilization; sisäinen suorituskyky)
 - komponentin i käyttöaste
 - komponentin i jononpituus
 - komponentin i "läpäisy aika"
- Kapasiteetti
 - järjestelmän kapasiteetti
 - komponentin i kapasiteetti

Suorituskyvystä ei voi puhua määrittelemättä työkuormaa

Vastausaikaperhe



Käyttöastesuureet

- Käyttöastesuureet ovat tutkijan mittoja
- Kasvun logiikka
 - kuorma kasvaa
 - laitteiden kuormitukset kasvavat (*eri tavoin*)
 - työjonoit laitteilla kasvavat (*ylilineaarisesti!!*)
 - töiden läpimenoajat kasvavat
- Tutkijan kysymykset
 - miten kuorman kasvu vaikuttaa eri laitteisiin
 - kuormitukseen, läpimenoaikaan, läpäisyyn
- Managerin uteliaisuus: kuinka paljon kapasiteettia tarvitaan?
- Käyttäjän valinta: onko yhteisten resurssien käytölle vaihtoehtoja?

Muita "suorituskykyysuureita"

Palvelun onnistuminen	Pyynnön tulos	Suorituskykyysuure
Palvelu tehty	Tulos: ok	Vastausaika Läpäisy Kuormitus
	Tulos: virhe i	Virheen i todennäköisyys Virheiden i väliaika
Palvelua ei voitu tehdä	Häiriö j	Häiriön j kesto aika Häiriöiden j väliaika

Kuinka hyvin järjestelmä pystyy palvelemaan
 - on saatavilla
 - suorittaa annetun tehtävän ja oikein

Heterogeenisyys

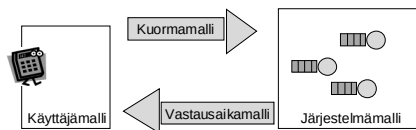
- Kuormamalli: joukko oleellisesti erilaisia töitä
 - kokonaiskäyttäytyminen ("keskiarvo"?)
 => työluokkakohittaiset suorituskykyarvot
- Epähomogeeninen aikakäyttäytyminen
 - "ruuhka-aika", "normaalkuorma", "hiljainen kausi"
 => mallinna erikseen

HUOMAA: mediaani "korjaa" jakauman vinouden, mutta se ei korjaa jakauman monihuippuisuutta!

Suorituskykyyn vaikuttavat tekijät

- Laitteiston komponenttien kapasiteetit
(kuorma & kapasiteetti => millainen kuormitus missäkin?)
- Ohjelmiston tehokkuus
 - "algoritmi" – ja sen toteutus
 - ohjelmistoarkkitehtuuri, erityisesti hajautus ja rinnakkaisuus
 - suorituskyky pullonkaulassa
- Työkuorma (workload)
- Resurssien hallinta
 - politiikka (allokointi, skedulointi)
 - viritys (parametrien asetukset)

Mistä puhutaan, kun puhutaan suorituskyvystä?



Käyttäjä => työkuorma
Työkuorma => resurssien kuormitus
Resurssien kuormitus => vastausaika

Malli: oleellisen täsmällinen kuvaus

Analyttinen malli



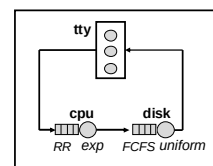
- Muodostetaan järjestelmää kuvaava matemaattinen malli
 - jonomalli
 - jonoverkkomalli
 - määritellyt: yleensä todennäköisyysjakaumia
- Määritetään tai määritellään mallin parametrien arvot
- Lasketaan suorituskykyseurainten arvot
- Menetelmäpaletti laaja
 - koulumatematiikka
 - jonoteoria
 - tasapainoyhtälönselitys (jne.)
- Mallin validiteetti? Tulosten validiteetin rajat?

Menetelmän valinta

Kriteeri	Mittaus	Simulointi	Analyttinen
Vaihe	proton	milloin vain	milloin vain
Ajan tarve	jälkeen vaihtelee	kohtalainen	pieni
Välineistö	"mittaristo"	ohjelmointikie- li (*)	analyttikko
Tarkkuus	vaihtelee	kohtalainen	karkea
Entä jos	vaikeaa	onnistuu	helppoa
Kustannus	korkea	kohtalainen	pieni
"Myytävyys"	korkea	kohtalainen	hankala

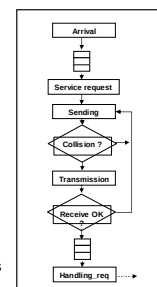
(*) ohjelmistotukea on olemassa
(**) aina mahdollista: tulokset ovat täysin virheellisiä

Mallin rakenne



Abstrakti matemaattinen rakenne
(esim. laskentaa varten)

Yksityiskohtainen toimintakuvaus
(esim. simulointia varten)



Mallin tarkkuus

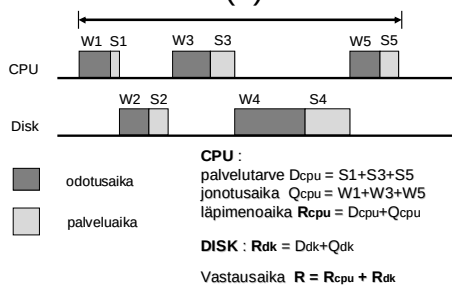
- Yksinkertainen malli => yleiskuva käyttäytymisestä
(analyttinen malli, *simulointi*)
 - Yksityiskohtainen malli => selityksiä käyttäytymiselle
(analyttinen malli, *simulointi*)
 - Mallinnusprosessi
 - aloita yksinkertaisella mallilla (ks. Occamin partaveitsi)
 - jos tulokset eivät tyydyttäviä, lisää yksityiskohtaisuutta: oleelliseksi arvioitavia selittäjiä
 - jatka, kunnes tietämyksen taso on riittävä (tai kustannukset kasvavat liian suuriksi)
- Tarkkuuden lisäyksen hinta:

Missä aika kuluu ...

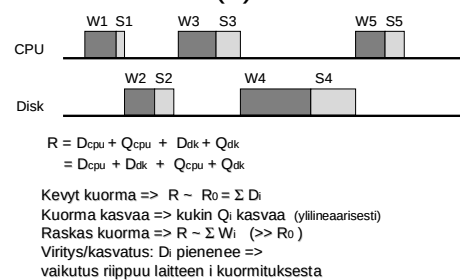
Network time		Server site time					
Latency	Transmission time	Service time			Queuing time		
		CPUs	Disks	LANs	CPUs	Disks	LANs

Latency: protokollan yleishallinto
Yksinkertaistus: verkon jonotuksia ei ole huomioitu

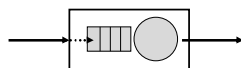
Palvelijan vastausajan synty (1)



Palvelijan vastausajan synty (2)



Asiakkaat jonossa



Operationaalinen jonomalli: kaikki suureet ovat mitattavissa

$S = B/C$ keskimääräinen palveluaika
 $U = B/T$ käyttöaste

Käyttöaste $U = \lambda S$

T mittausaika
A saapuvien lkm
C poistuvien lkm
B "busy time"

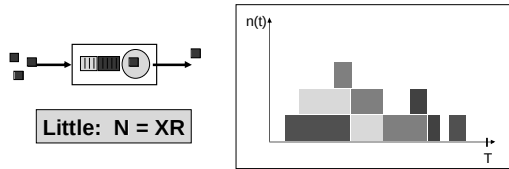
Esimerkki

Verkkosegmentti

- läpäisy mittausaikana 1000 pakettia/s
- paketin siirtoaika 0.15 ms
- mikä on segmentin käyttöaste?
- $U = XS \Rightarrow U = 1000 \times 0.00015 = 0.15 = 15\%$

- Oletus operationaalisissa malleissa
 - mitattu ajanjakso on "tyypillinen"
 - järjestelmä on "tasapainotilassa" (A~C)

Littlen laki



Little: $N = XR$

W = kokonaisoleskeluaika järjestelmässä (tumma alue)

$R = W / C$ keskimääräinen vastausaika

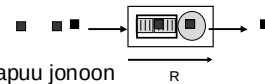
$N = W / T$ keskimääräinen jonopituus (*)

$X = C / T$ läpäisy

(*) Huom: "jononpituus" sisältää myös palveluttavan

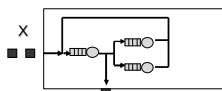
Littlen laki: "perustelu"

Saapumistiheys X



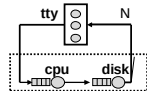
- Asiakas saapuu jonoon
 - jonon pituuden odotusarvo on N
 - Asiakkaan "läpimenoaika"
 - läpimenoajan odotusarvo on R
 - tänä aikana jonoon saapuu uusia asiakkaita XR
 - Asiakas poistuu jonosta
 - jälkeen jäävän jonon pituuden odotusarvo on N
- $\Rightarrow N = XR$

Avoin ja suljettu järjestelmä



Avoin järjestelmä

- "(ääretön) populaatio"
- asiakkaiden lkm vaihtelee
- saapumistiheys = poistumistiheys
- kysymys:
 - vastausajan käyttäytyminen



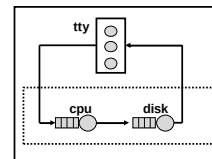
Suljettu järjestelmä

- kiinteä asiakasjoukko
- kysymys:
 - vastausajan käyttäytyminen
 - "vuon" nopeus
 - (käyttöasteet?)

Vastausaikalaki (response time law)

M päätettä
 Z miettimisaika

R vastausaika
 X suoritusteho
"kyselyä/aikayksikkö"

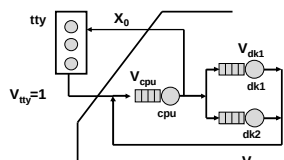


Little: $M = X(Z + R)$

$\Rightarrow R = M/X - Z$

Esim: $M = 12$, $X = 0.5$ vast/s, $Z = 20$ s
 $\Rightarrow R = 4$ s

Pakkovuolaki (forced flow law)



- T mittausaika
- C_0 vastausten lkm
- X_0 vastausaika
- C_i töitä laitteella i

Työn käyntisuhde (visit ratio) laitteella i : $V_i = C_i / C_0$

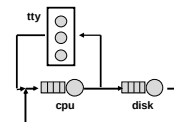
$X_i / X_0 = C_i / C_0 = V_i$

$\Rightarrow X_i = V_i X_0$

$X_0 \sim$ vuon nopeus; $\{V_i\} \sim$ vuon jakautuminen verkkoon

Esimerkki: kyselyjärjestelmä

- Kysely
 - $V_{sk} = 20$
 - $S_{sk} = 25$ ms



Oletettu käyttötilanne

- aktiivisia käyttäjiä $M = 25$
- miettimisaika $Z = 18$ s
- levyn käyttöaste $U_{sk} = 0.5$

$R = M/X_0 - Z$

$X_0 = ?$

$X_i = V_i X_0 \Rightarrow X_0 = X_i / V_i$

$X_i = U_i / S_i \Rightarrow X_0 = U_i / (V_i S_i)$

$X_0 = 0.5 / (20 \cdot 0.025) = 1$

$R = M/X_0 - Z$ Määritelmä: $V_i = D_i$ (demand)

Moniluokkainen työkuorma

- Oletus edellä: työt ovat "tilastollisesti samanlaisia" (kuorman töitä mallinnetaan "keskiarvotyöllä").
- Moniluokkainen kuormamalli: kuorma muodostuu keskenään erilaisista luokista, kunkin luokan työt ovat sisäisesti "samanlaisia"
- Esimerkkejä
 - pankkiautomaatti: saldokyselyt / setelinostot
 - tietoliikenne: kuittaukset / kyselyt / tekstitiedoston siirto / videokuvan siirto
 - www-selaus: kyselyt / vastaukset

Syitä moniluokkaisuudelle

Perussyitä: heterogeenisyyden hallinta

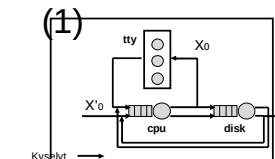
- Töiden erilaiset resurssinkäyttöprofiilit
- Kuorman saapumiskäyttäytymisen erilaisuus: avoin tai suljettu luokka
- Erilaisilla työtyypeillä erilaiset suorituskykyvaatimukset (kysely / päivitys / ylläpito)
- Erilaisilla työtyypeillä erilaiset kasvunesteet

Heterogeenisyys ajan yli

- kukin "homogeeninen" periodi mallinnetaan erikseen (esimerkiksi vuorokaudenajat: ruuhka / normaali / kevyt)

Esimerkki: kaksi työluokkaa

- Kyselyjärjestelmä
 - päätteitä $M = 40$
 - miettimisaika $Z = 15s$
 - vastausaika $R = 5s$
- Levyn käyttö
 - $S_{dk} = 40 ms$, $U_{dk} = 0.9$
 - kyselyt: $V_{dk} = 10$
 - erätyöt: $V'_{dk} = 5$

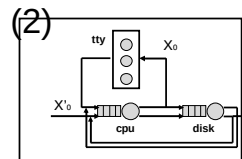


- Erätyöt:
- $X'_0 = ?$
 - jos $X'_0 > 3$ niin $R \rightarrow ?$

Lähde: Denning, Buzen, The Operational Analysis of ...; ACM CS 10 (3), Sept. 1978

Esimerkki: kaksi työluokkaa

- Kyselyjärjestelmä
 - päätteitä $M = 40$
 - miettimisaika $Z = 15s$
 - vastausaika $R = 5s$
- Levyn käyttö
 - $S_{dk} = 40 ms$, $U_{dk} = 0.9$
 - kyselyt: $V_{dk} = 10$
 - erätyöt: $V'_{dk} = 5$



- Levy**
- kyselyjärjestelmä $X_0 = M/(Z+R) = 40/20 = 2 \text{ kys/s}$
 - kyselyjen läpäisy $X_{dk} = V_{dk}X_0 = 20 \text{ noutoa/s}$
 - kokonaisläpäisy $X_{dk} + X'_{dk} = U_{dk}/S_{dk} = 22.5 \text{ noutoa/s}$
 - erätöiden läpäisy $X'_{dk} = 22.5 - 20 = 2.5 \text{ noutoa/s}$

Koko järjestelmä: erätöiden läpäisy $X'_0 = X'_{dk}/V'_{dk} = 0.5 \text{ työtä/s}$

- Pelkistetyt välineet antavat toisinaan virheellisen tuloksen
 - Otettava huomioon erilaisia todennäköisyysjakaumia ja skedulointitapoja
 - Teoriaa: suorituskykyanalyysin kurssi kurssikirjojeen
 - Erilaisia arviointityökaluja saatavilla
 - HJ-kurssilla "ad hoc" –arvioita – peruskaavojen taso riittää tuomaan esiin oikeat kysymyskohdat