Scheduling and Quality of Service

Manuel P. Ricardo

Faculdade de Engenharia da Universidade do Porto

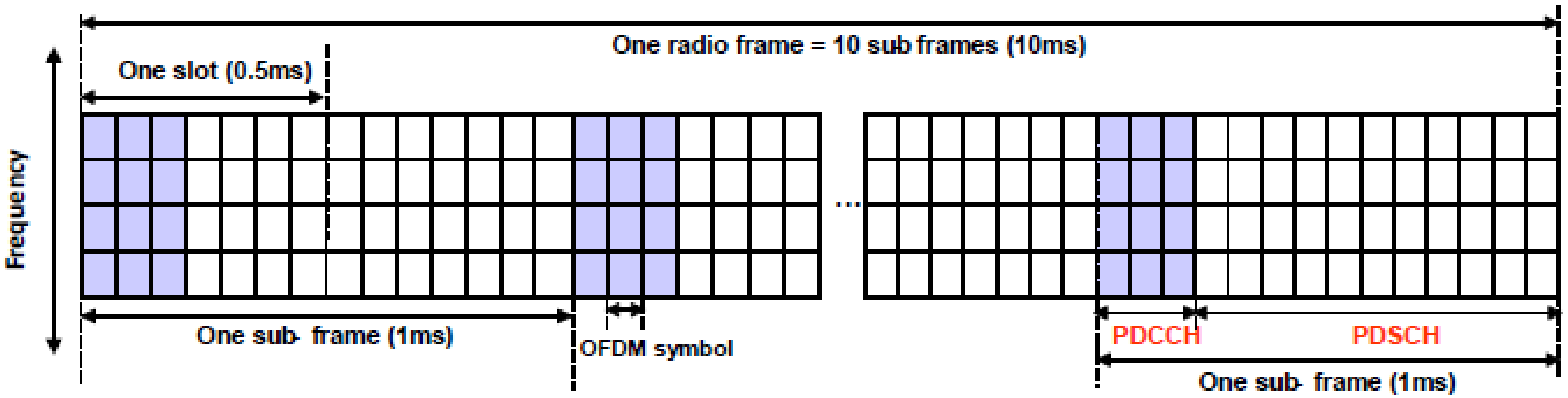
Scheduling

Radio Resource Allocation

- Contention-based access protocols
  - Distributed resource usage and allocation
  - Low efficiency when many terminals access the network (many collisions)
- Reservation-based access protocol with centralized scheduling
  - Commonly used in cellular networks
  - High efficiency and flexibility

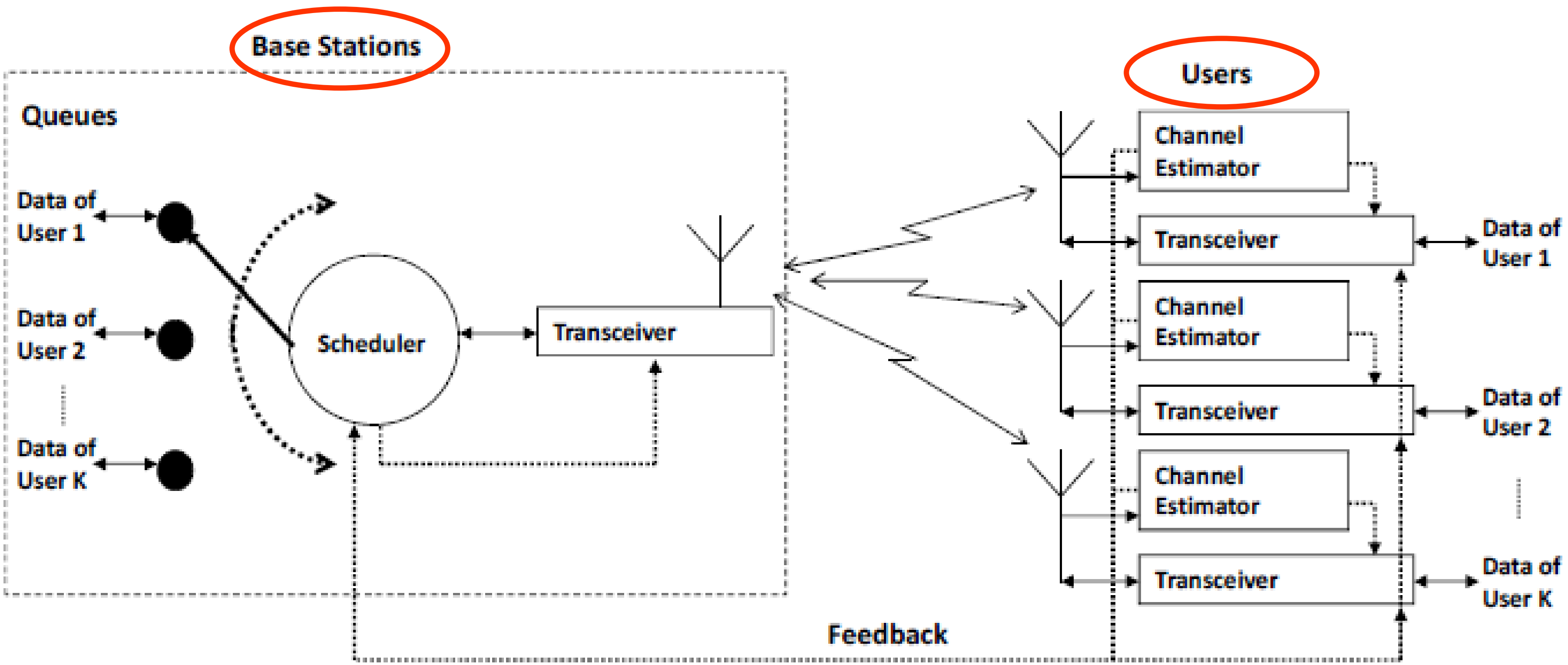
A Reservation-based Protocol (LTE, NB-IoT)

- Physical Downlink Control Channel (PDCCH)
  - » conveys control information for each user
- Physical Downlink Shared Channel (PDSCH)
  - » multiplex the data of all terminals
  - » Each user receives on a unique set of OFDM symbols and frequency blocks
- Reservation phase: PDCCH | Data phase: PDSCH



Reservation Phase

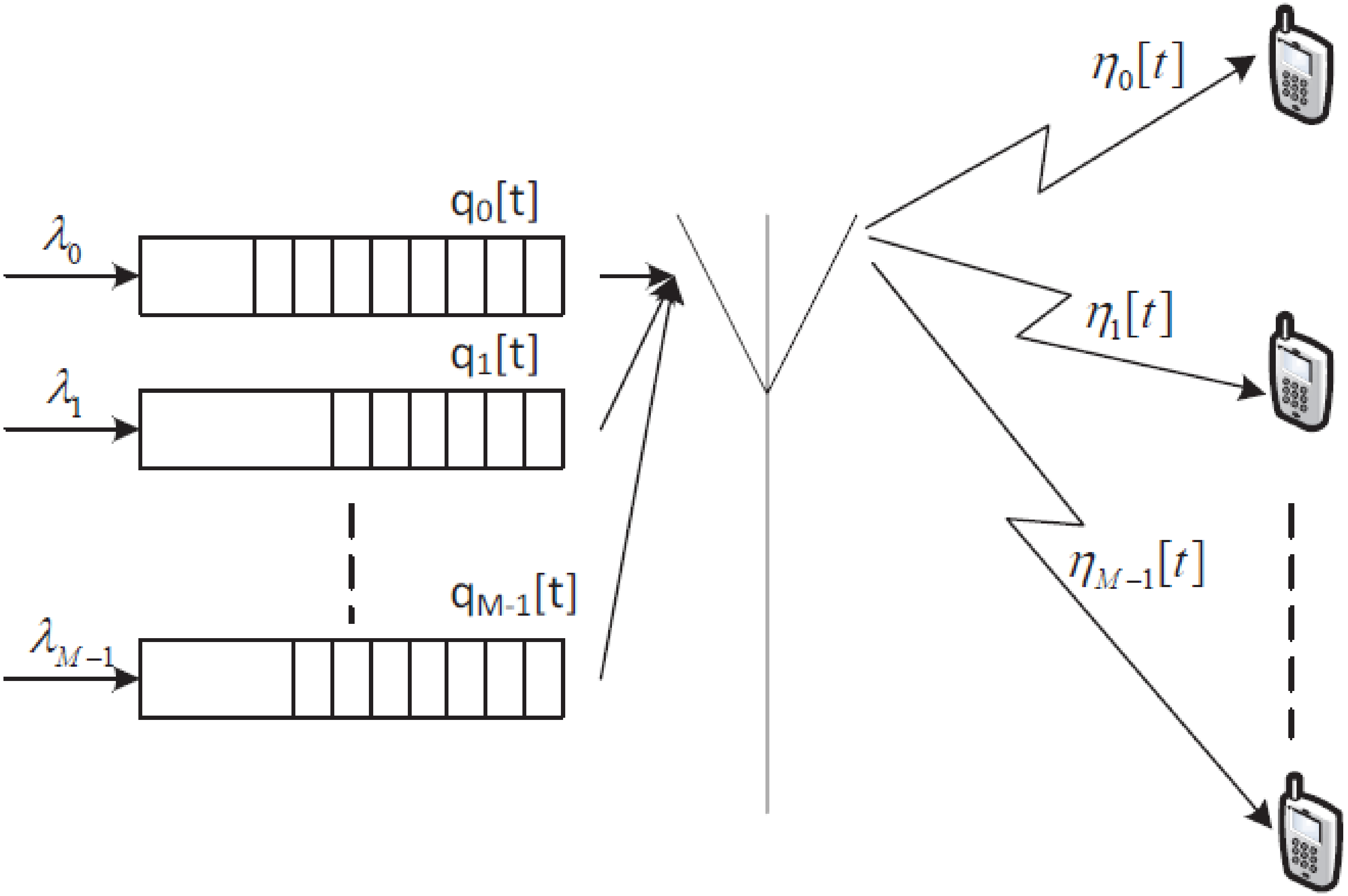
- Estimate channel → Provide feedback → Schedule



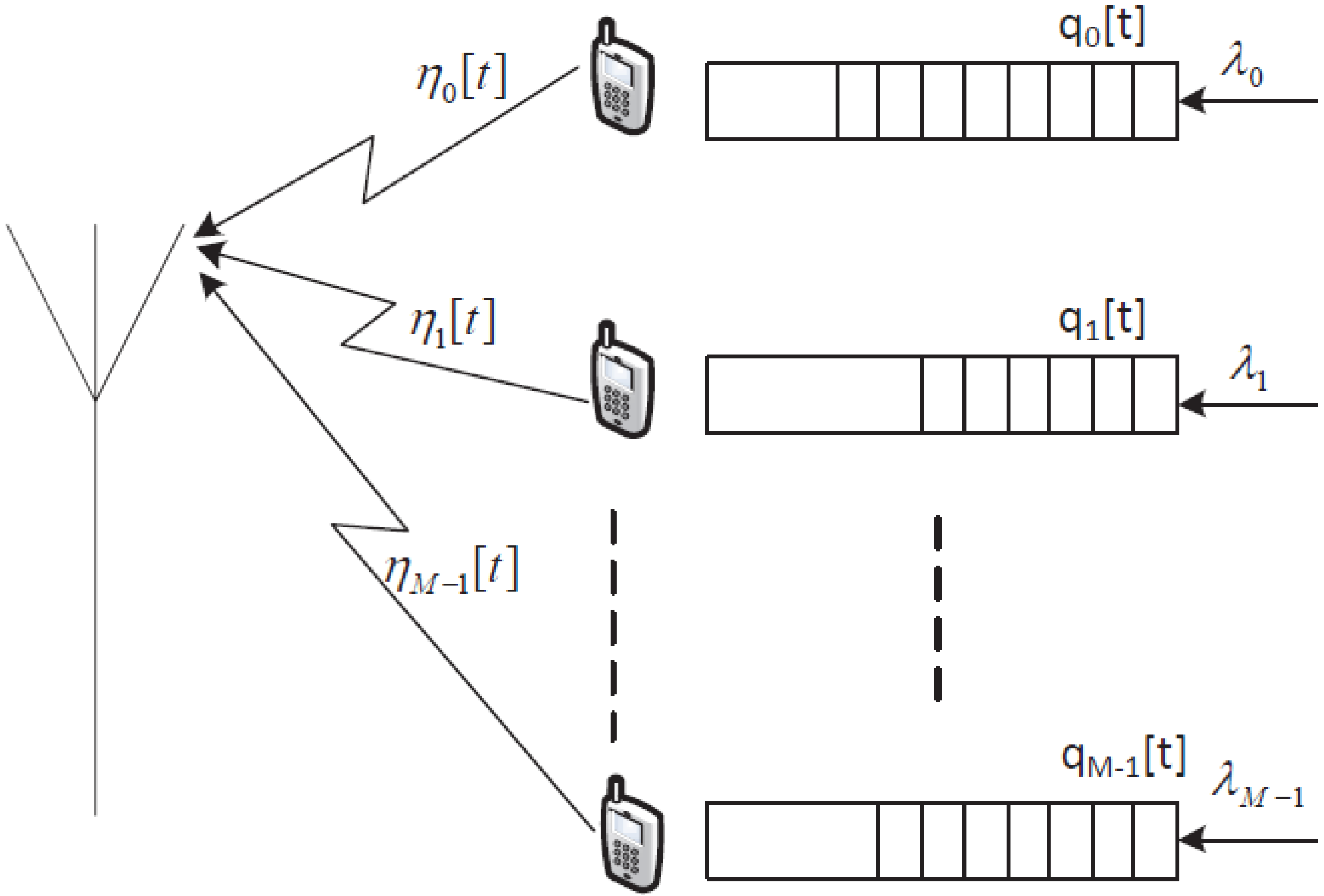
Packet Scheduling Algorithms – Simple Model

- M users served on a single channel
- TDMA: one user in one slot
- Each user has a buffer (to store the information to be transmitted)

Downlink Scheduling



Uplink Scheduling



Queue Modeling

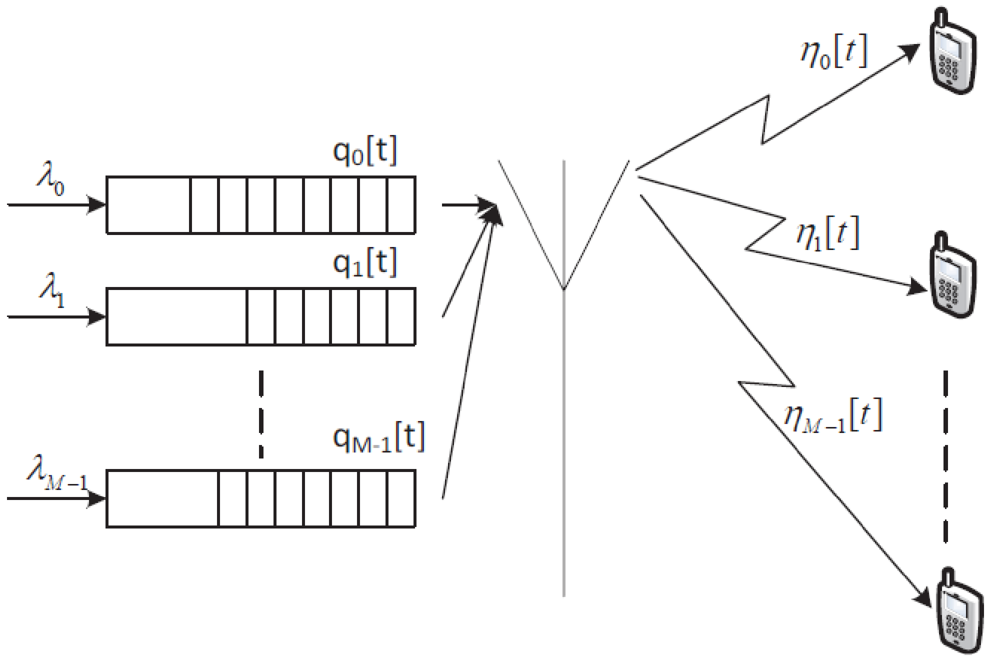
- At timeslot t, the queue of user i changes as follows

$$q_i[t+1] = q_i[t] + \delta_i[t] - \eta_i[t]$$

Number of bits in the queue

Number of new bits arriving

Number of bits scheduled to transmit, determined by scheduling.



Round-Robin (RR) Scheduling

- Users are scheduled in a round robin (cyclic order)
- $i[t]$: user scheduled at time t
- RR scheduler: $i[t+1] = i[t] + 1 \pmod{M}$
- Fair: all users scheduled the same amount of resources

Max Throughput Scheduling

- Objective: Maximize total network throughput

- If user ***i*** is scheduled, the expected data rate is

$$\hat{r}_i[t] = \frac{\hat{\eta}_i[t]}{T_s}$$

Expected number of bits that can be successfully delivered

Slot length

- The total expected network throughput is

$$\hat{r}[t] = \sum_{i=0}^{M-1} \hat{r}_i[t] I(i)$$

I(i) : Scheduling indicator:
1 scheduled, 0 otherwise.

Max Throughput Scheduling

- Schedules the user with the highest expected data $\hat{r}_i[t]$

$$\hat{r}_i[t] = W \log_2\left(1 + \frac{\Gamma_i[t]}{\theta}\right)$$

  - » W : channel bandwidth
  - » $\Gamma_i[t]$: SINR at time t given the allocated power
  - » θ : SINR gap (gap between channel capacity and practical coding and modulation scheme)

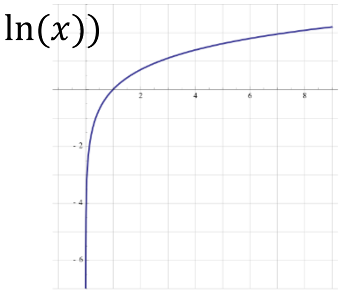
  ➔ Schedule the user with max SINR

- Drawbacks
  - » Unfair
  - » Coverage limitation
  - » Some users may never be served

Proportional Fair (PF) Scheduling

- Balance competing interests
  - » network throughput vs minimum service level

- Objective: Maximize $\boxed{\sum_{i=0}^{M-1} \ln S_i}$



- S_i : long-run throughput for user ***i*** , predicted by
  - » $$\hat{S}_i[t] = (1 - \frac{1}{\tau})S_i[t-1] + \frac{1}{\tau}\hat{r}_i[t]I(i)$$
  - » where $\tau >> 1$ is a constant defined by the scheduler

➔ Schedule the user with the highest $\boxed{\dfrac{\hat{r}_i[t]}{S_i[t-1]}}$

Max-Min Scheduling

- Objective: Maximize the minimum user throughput $\boxed{\max_i \min S_i}$

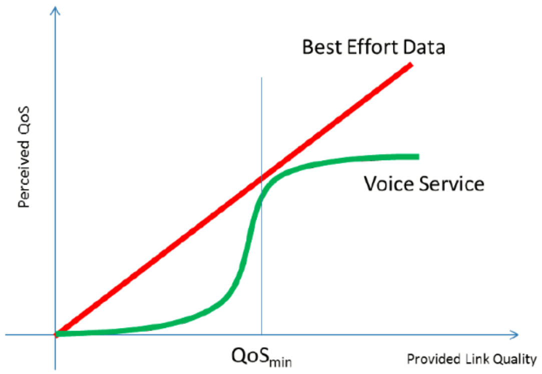
- A scheduling result is max-min fair if and only if a further increase of throughput of one user will result in the decrease of a user with a smaller throughput

$$\hat{S}_i[t] = (1 - \frac{1}{\tau})S_i[t-1] + \frac{1}{\tau}\hat{r}_i[t]I(i) \qquad \max \min_i (1 - \frac{1}{\tau})S_i[t-1] + \frac{1}{\tau}\hat{r}_i[t]I(i).$$

➔ Schedule the user with the minimum $\boxed{(1 - \frac{1}{\tau})S_i[t-1]}$

the one with the smallest throughput at time t-1

Max Utility Scheduling



- Previous schedulers do not consider QoS

- Utility-based scheduling
  - » Utility quantifies the satisfaction of each user
  - » Model the users QoS perception

- Objective: Maximize the sum utility of all users

  (total network satisfaction)

- Utility functions: model how user perceives services

Max Utility Scheduling

$$\max \sum_{i=0}^{M-1} U_i(S_i)$$

- Different utility functions define different fairness and efficiency

- Max Throughput Scheduling (highest efficiency): $U(S) = S$

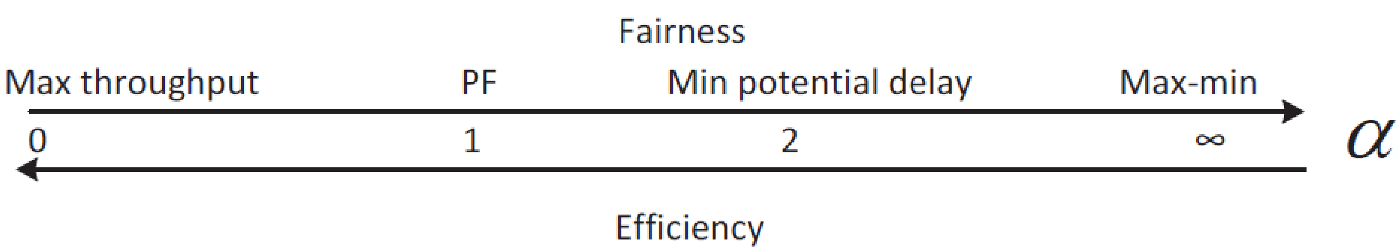
- Proportional Fair Scheduling : $U(S) = \ln(S)$

Max Utility Scheduling - Alpha Fair Utility

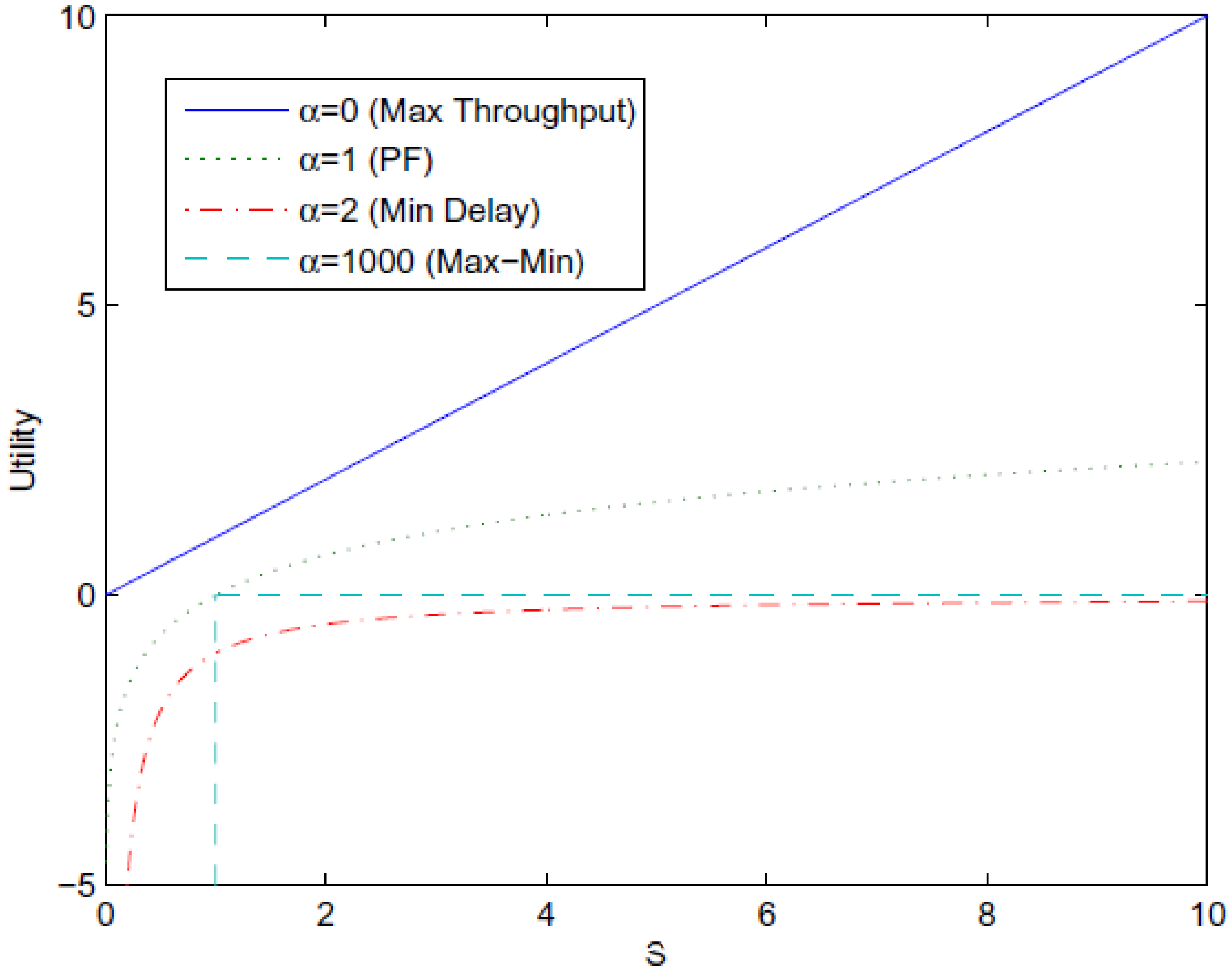
- More generic definition: α fair scheduling

$$U_\alpha(S) = \begin{cases} \frac{S^{1-\alpha}}{1-\alpha} & \alpha \geq 0 \text{ and } \neq 1 \\ \ln(S) & \alpha = 1. \end{cases}$$

- α measures how fair the scheduling result is
  - » 0 : Max throughput
  - » 1 : Proportional fair
  - » 2 : equivalent to min $\sum_{i=0}^{M-1} \frac{1}{S_i}$ (minimize the total delay, sec/bit)
  - » Infinity: Most fair, max-min scheduler

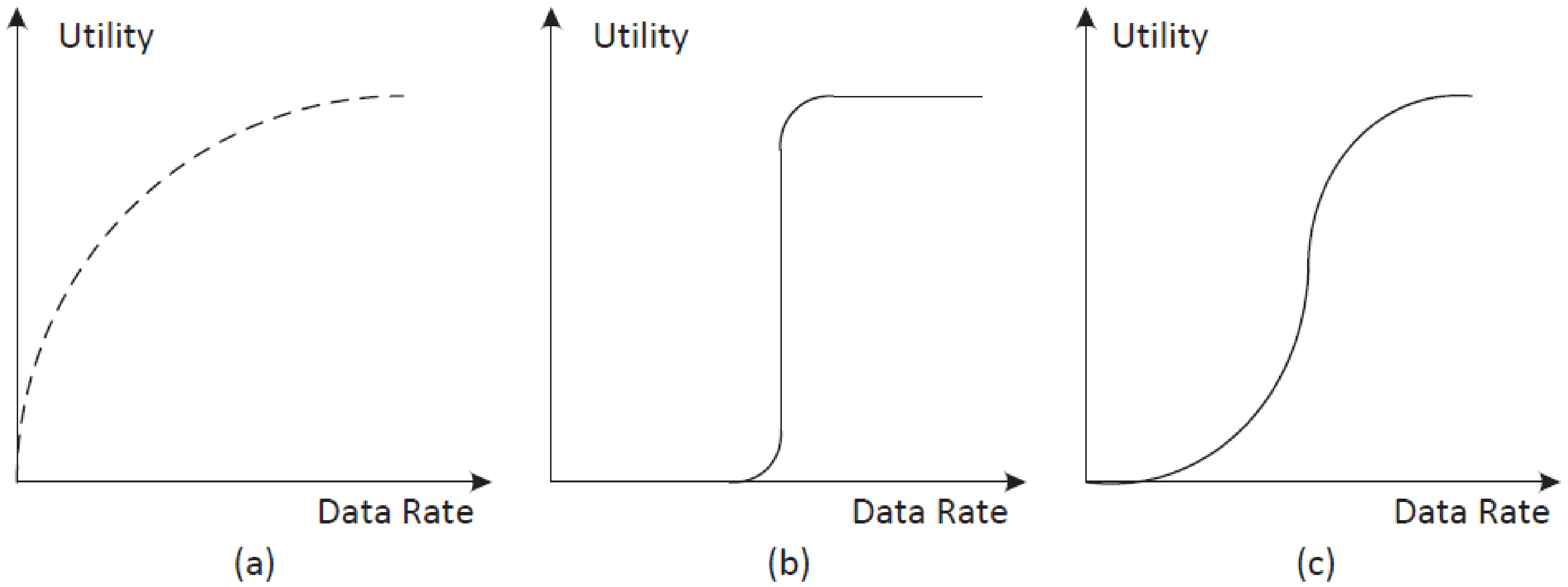


Alpha Fair Utility



Utility Functions with QoS Consideration

- Determined based on traffic characteristics



(a) best effort

(b) real time with tight delay requirement

(c) real time with loose delay requirement

How to Schedule Users in Utility Scheduling

$$\max \sum_{i=0}^{M-1} U_i(\hat{S}_i[t]) = \max \sum_{i=0}^{M-1} U_i\left((1-\frac{1}{\tau})S_i[t-1] + \frac{1}{\tau}\hat{r}_i[t]I(i)\right)$$

Since $(1-\frac{1}{\tau})S_i[t-1] >> \frac{1}{\tau}\hat{r}_i[t]I(i)$

We have

$$U_i\left((1-\frac{1}{\tau})S_i[t-1] + \frac{1}{\tau}\hat{r}_i[t]I(i)\right)$$
$$\approx U_i\left((1-\frac{1}{\tau})S_i[t-1]\right) + U_i'\left((1-\frac{1}{\tau})S_i[t-1]\right)\frac{1}{\tau}\hat{r}_i[t]I(i)$$

- Since $U_i\left((1-\frac{1}{\tau})S_i[t-1]\right)$ is fixed at time t
- Equivalent objective: $\max \sum_{i=0}^{M-1} U_i'\left((1-\frac{1}{\tau})S_i[t-1]\right)\frac{1}{\tau}\hat{r}_i[t]I(i)$

➔ Schedule the user with the largest: $U_i'\left((1-\frac{1}{\tau})S_i[t-1]\right)\hat{r}_i[t]$

or $\boxed{U_i'(S_i[t-1])\,\hat{r}_i[t]}$ (because τ >> 1 ➔ 1/τ → 0)

Performance Comparison - Example

- Users requesting different types of services
- Ten users randomly positioned in the cell

Traffic	Type	Basic rate requirement
VoIP	Real Time	102 Kbps
Video Streaming	Real Time	580 Kbps
High Rate File Download	Best effort	1.74 Mbps

Schedulers	Average throughput (Mbps)	Outage probability (%)
Max-SINR	17.4	5.8
Round Robin	9.1	2.9
Proportional	12.7	2.0
Utility	9.5	0

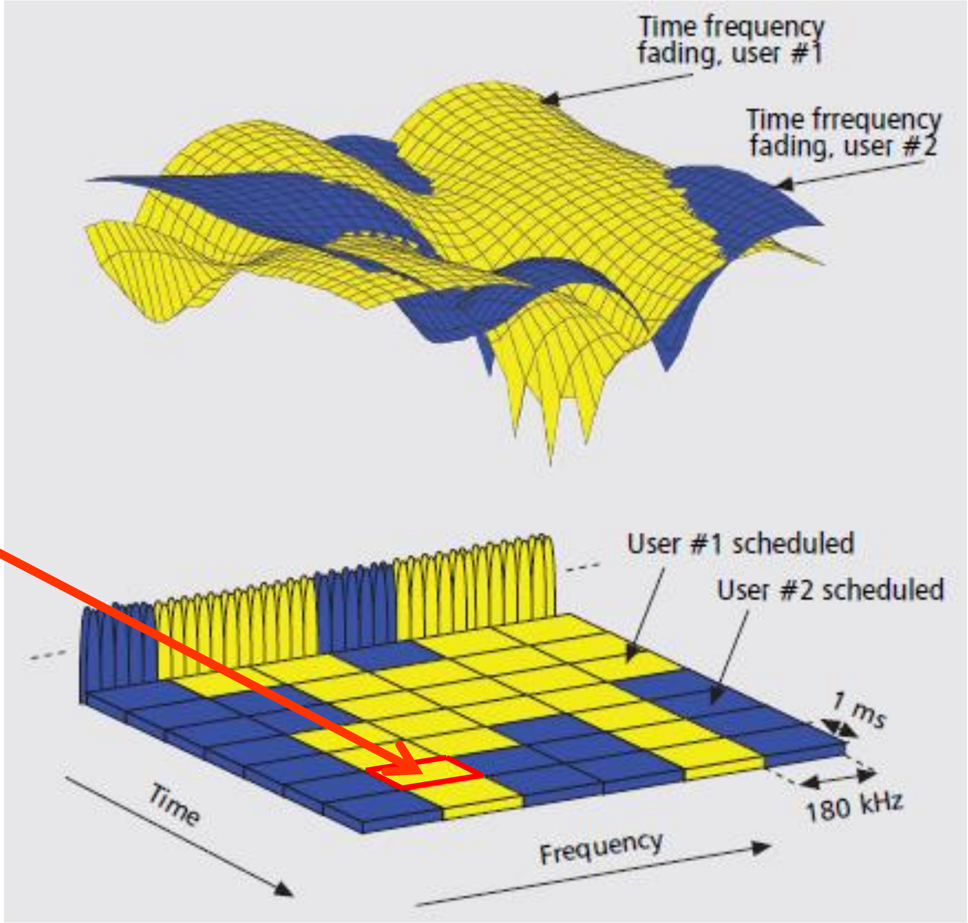
Scheduling in OFDMA Systems

- One more dimension of resources
  - → sub-carrier allocation

- Different users
  - » experience independent wireless channels
  - » subcarriers have different channels gains

Example - LTE Radio Resources

- LTE uses **OFDMA**
- Time x Frequency space
- Radio Block ***RB***
  - » $T_{RB} \times B_{RB} = 1ms \times 180\,kHz$
  - » Schedulable resource unit
- Blocks are allocated to flows



Max-Throughput OFDMA Scheduling

- Let us assume OFDMA uses Max-Throughput Scheduling

- Throughput of user ***i*** in $\boldsymbol{t^{th}}$ OFDM slot

$$R_i[t] = \sum_j W \log_2 \left(1 + \frac{p_{ij}[t]\gamma_{ij}[t]}{\theta} \right) I(i, j)$$

  - » j : j-th subcarrier
  - » W : subcarrier bandwidth
  - » $p_{ij}[t]$: power allocation on the j-th subcarrier of terminal i
  - » $\gamma_{ij}[t]$: ratio gain - interference, SINR = $p_{ij}[t]$ * $\gamma_{ij}[t]$
  - » I(i, j) : 1 if the j-th subcarrier is assigned to terminal ***i*** and 0 otherwise

Max-Throughput OFDMA Scheduling

- Each subcarrier assigned to a single user

$$\sum_i I(i, j) = 1, \forall j.$$

- Overall network throughput

$$R[t] = \sum_{i=1}^{M} R_i[t] = \sum_i \sum_j W \log_2\left(1 + \frac{p_{ij}[t]\gamma_{ij}[t]}{\theta}\right) I(i, j)$$

Example - Priority Set Scheduler

- Time Domain scheduler, in frame t
  - » Classifies flows in 2 sets, considering past average throughput $\boldsymbol{S_i[t-1]}$

$$\text{flow } i \in H \quad \text{if} \quad S_i[t-1] < S_i, \quad \text{ordered by } \frac{1}{S_i[t-1]}$$

$$\text{flow } i \in L \quad \text{if} \quad S_i[t-1] \geq S_i, \quad \text{ordered by} \longrightarrow \frac{\hat{r}_i[t]}{S_i[t-1]}$$

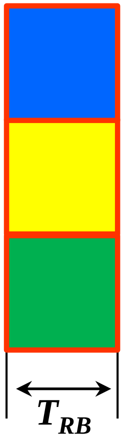
  - » Selects M flows. Flows in H have priority over flows in L

- FD scheduler, associates blocks to the M flows using a *Proportional Fair* technique

Block k is associated to flow $i \in M$ such that

$$i_k[t] = \arg\max_{j=1,\cdots,M} \left(\frac{\hat{r}_j[k,t]}{S_j[t-1]} \right)$$

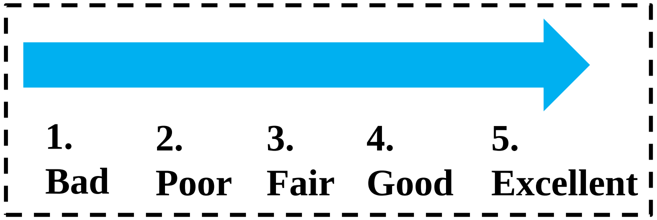
$\hat{r}_j[k,t]$ - estimated throughput for flow j on RB k



Quality of Service

Quality of Experience

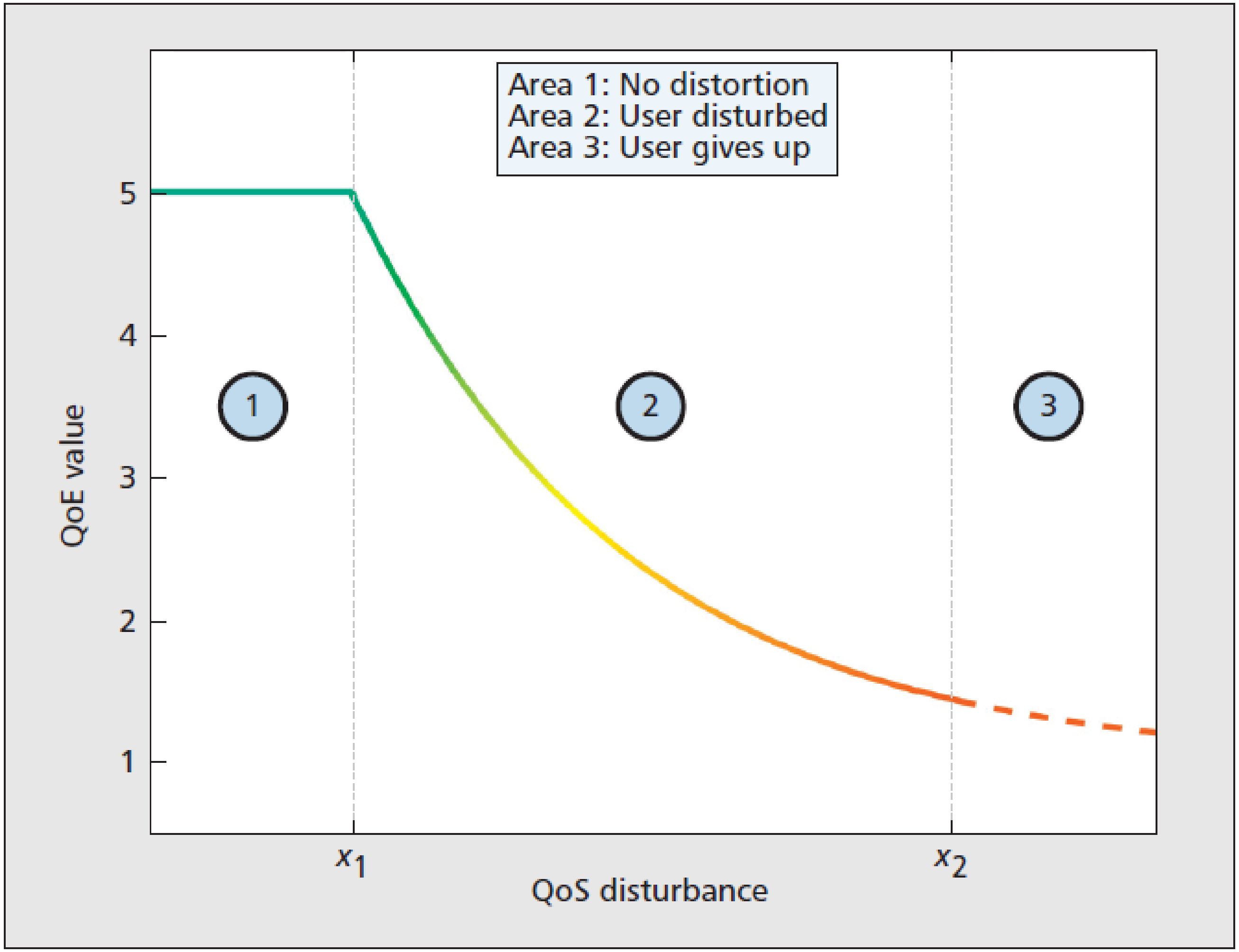
- Evaluated by panels
  - Mean Opinion Score (MOS)



- Measured objectively
  - Selection of Key Quality Indicators or Quality of Service parameters

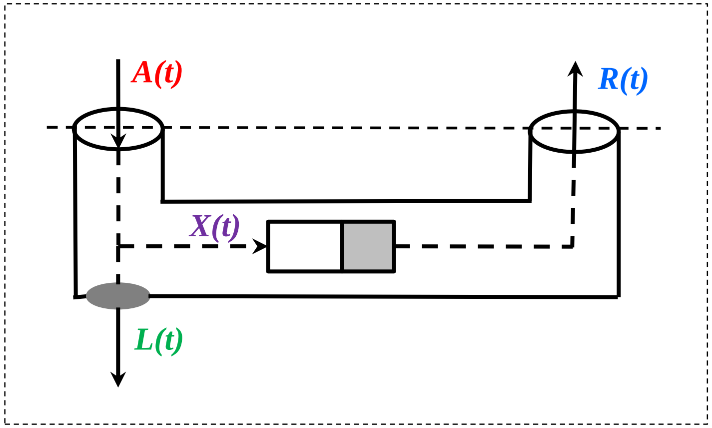
Layer	**Indicators / Parameters**	
	VoIP	**Video streaming**
Application	R-factor, speech distortion, acoustic echo, SNR speech, latency speech	PEVQ, PSNR, frame rate, pixilization, video frame loss, lip-sync, contrast
Network (packet switched)	***Packet Delay, Packet Loss Ratio, Throughput***	

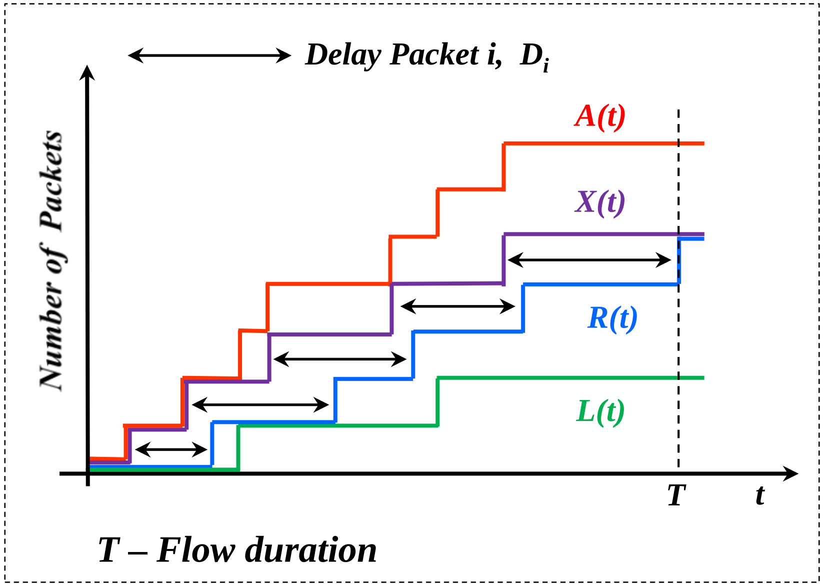
QoE vs QoS



Area 1: No distortion
Area 2: User disturbed
Area 3: User gives up
5
4
3
2
1
QoE value
1
2
3
x_1
x_2
QoS disturbance

QoS Parameters (network layer)



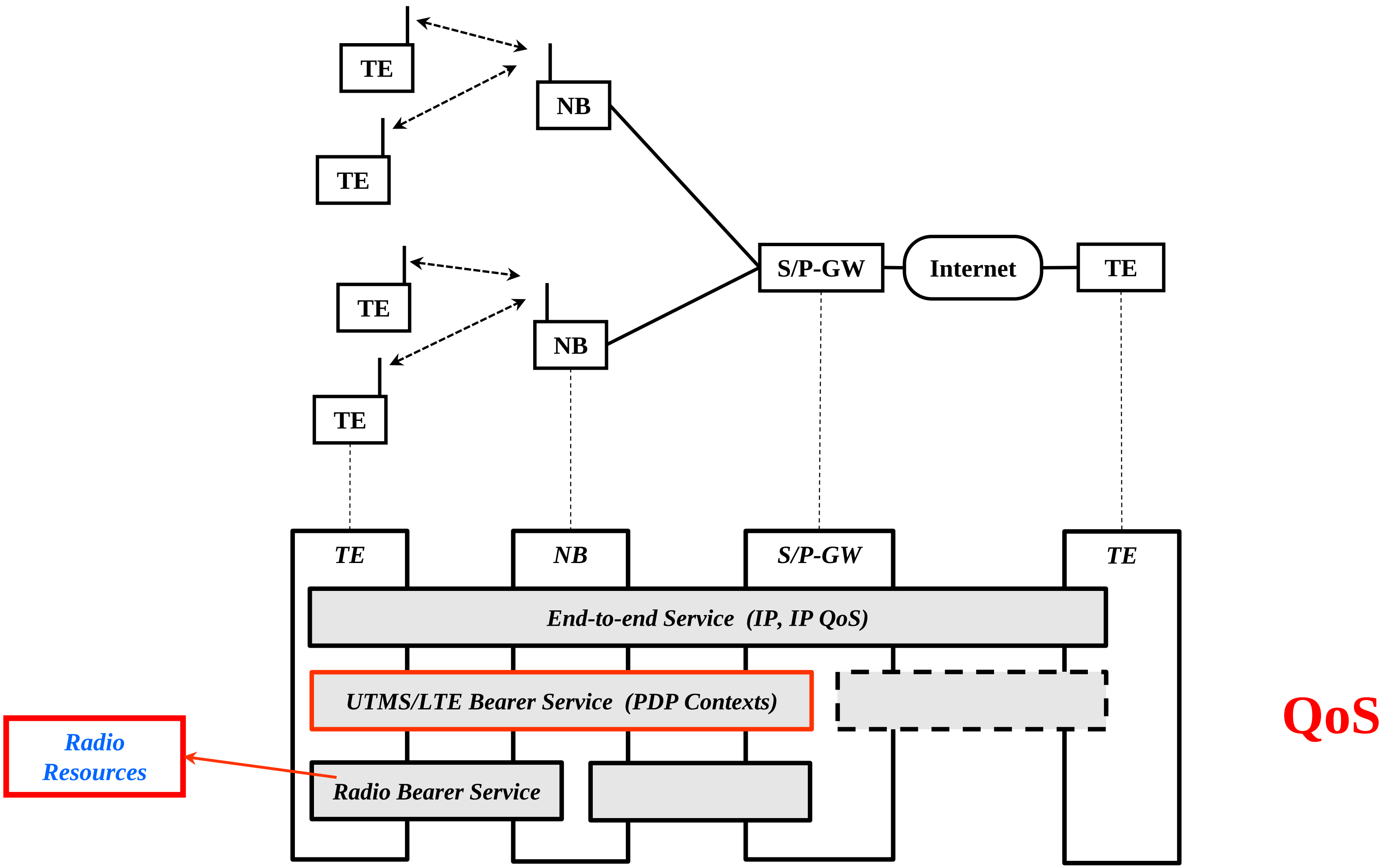


$$\overline{PacketDelay} = \frac{\sum_{i=1}^{R(T)} D_i}{R(T)}$$

$$PacketLoss\,Ratio = \frac{L(T)}{A(T)} = \frac{A(T) - R(T)}{A(T)}$$

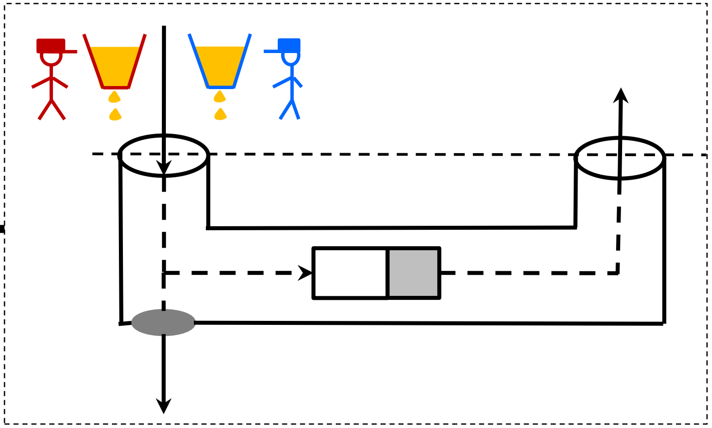
$$Throughput = \frac{R(T)}{T} \bullet \overline{L}, \quad \overline{L}: \text{average packet length}$$

*QoS Architecture**

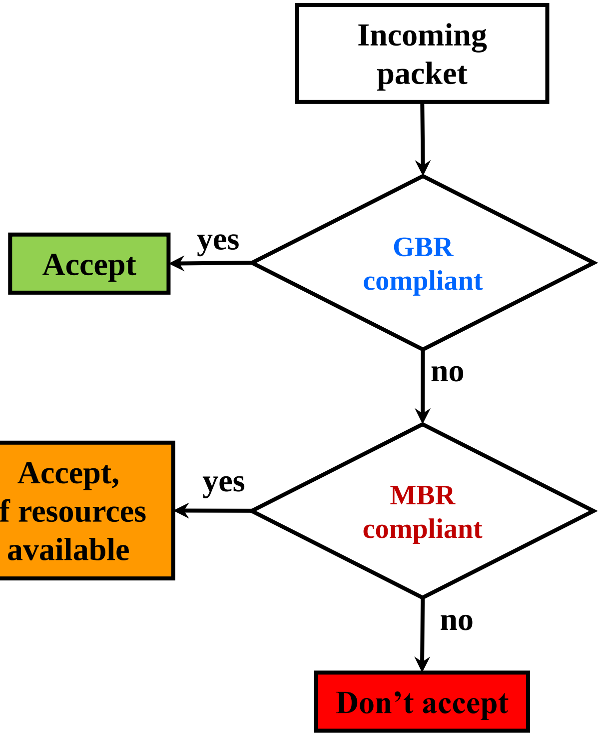


* *simplified representation*

Bearer Service - QoS Parameters



- *Packet Loss Ratio*
- *Transfer delay,* $P(D < TransferDelay) \geq 0{,}95$
  - Maximum delay guaranteed for 95% of packets
- *Guaranteed Bit Rate (GBR)*
  - Policed by token bucket ***(GBR, MaxSDUsize)***
- *Maximum Bit Rate (MBR)*
  - Policed by token bucket ***(MBR, MaxSDUsize)***



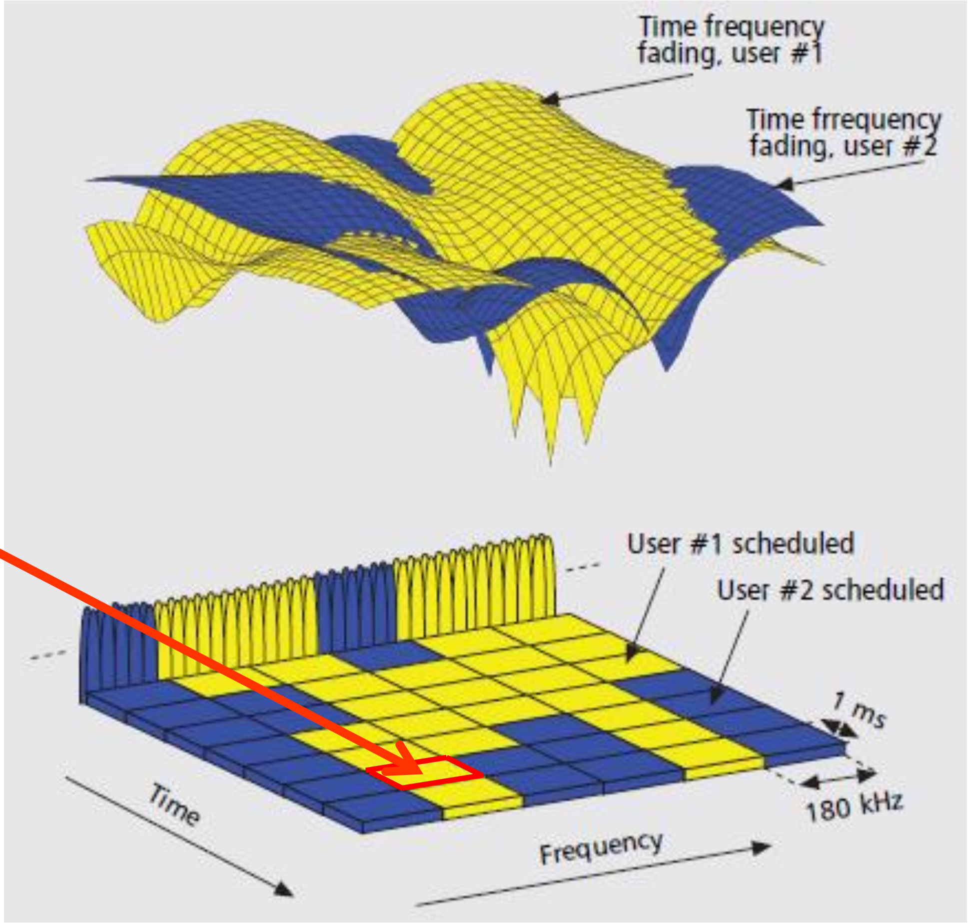
LTE Traffic Classes (4G)

Class	***Resource Type***	***Guaranteed Bit Rate (bit/s)***	***Maximum Bit Rate (bit/s)***	***Transfer Delay (ms)***	***Packet Loss Ratio***	***Priority***	***Application***
1	***GBR***	yes	yes	100	10^{-2}	2	Conversational voice
2				150	10^{-3}	4	Video streaming
3				50		3	Real-time gaming
4				300	10^{-6}	5	Buffered streaming
5	***Non-GBR***	no		100		1	IMS signalling
6				300		6	Video and TCP based apps
7				100	10^{-3}	7	Voice, video
8				300	10^{-6}	8	Video and TCP based apps

- GBR
  - » Guaranteed bitrate, low guaranteed delay, high packet loss ratio
- Non-GBR
  - » No bitrate guarantees, low packet loss ratio

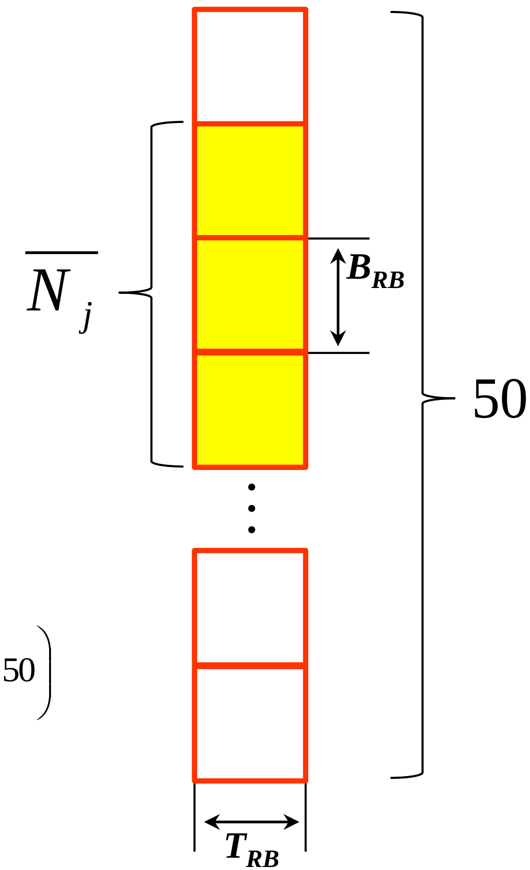
LTE Radio Resources

- LTE uses **OFDMA**
- Time x Frequency space
- Radio Block ***RB***
  - » $T_{RB} \times B_{RB} = 1ms \times 180\,kHz$
  - » Schedulable resource unit
- Blocks are allocated to flows



LTE Admission Control

- Arrival new flow ***j*** characterized by $\boldsymbol{R_j}$ *[bit/s]*
  $\overline{N}_j$ blocks/frame are required



- New flow ***j*** is admitted if

$$\boxed{\sum_{i=1}^{\#Admitted} \overline{N}_i + \overline{N}_j < N_{total}}$$

$$\left(N_{total} = \frac{B_{LTE}}{B_{RB}} = {}^{10\,MHz}\!/\!_{180\,kHz} \approx 50 \right)$$

» Information I_{RB} transported by a block

$$I_{RB}[bit] = k \bullet B_{RB} \bullet \log_2(1 + SNR_j) \bullet T_{RB}, \quad k < 1$$

» Average number of Blocks per frame

$$\overline{N}_j = \frac{R_j \bullet T_{RB}}{\overline{I}_{RB}}$$

UMTS Traffic Classes (3G)

Traffic Class	***Guaranteed Bit Rate (bit/s)***	***Maximum Bit Rate (bit/s)***	***Transfer Delay (ms)***	***Packet Loss Ratio***	***Priority***
Conversational	yes	yes	80 – max. value	10^{-2}, …, 10^{-5}	
Streaming			250 – max. value	10^{-1}, …, 10^{-5}	
Interactive	no			10^{-3}, …, 10^{-6}	1,2,3
Background					

- Conversational class
  - » Guaranteed bitrate, guaranteed delay (low), high packet loss ratio
- Streaming class
  - » Guaranteed bitrate, guaranteed delay (high), high packet loss ratio
- Interactive class
  - » Priorities (instead of guarantees), low packet loss ratio
- Background class
  - » Lowest priority, low packet loss ratio

UMTS Radio Resources

- UMTS uses ***Code Division Multiple Access***
- Simultaneous transmissions possible by using orthogonal codes
- Transmitted power causes also interference
- Transmitted powers have to be managed

UMTS Radio Resource Management

- Packet Loss Ratio
  - Used to define $\boldsymbol{P_t}$ PLR → BER → $\left(\frac{E_b}{N_0}\right)_j$ → SNIR → P_r → P_t
- Guaranteed Bit Rate
  - Used to control total interference (load) in cell (next slide)
- Transfer delay
  - Used to define ARQ operation mode (acknowledged, non-acknowledged)

QoS Class	***Conversational***	***Streaming***	***Interactive***	***Background***
Admission control	***Yes***		*No*	
Transport channels	***Dedicated (code)***		*Shared*	
Scheduling	***Non-scheduled***		*Scheduled by packet scheduler*	

UMTS Admission Control

- Arrival new flow ***j*** characterised by $\boldsymbol{R_j}$ *[bit/s]*
- New flow ***j*** is admitted if $\boxed{\eta + L_j < \eta_{\max}}$, where

$$\boxed{\eta = \sum_{i=1}^{Admitted} L_i}$$

$$\boxed{L_j = \frac{P_j}{I_t} = \frac{1}{1 + \frac{W}{\left(\frac{E_b}{N_0}\right)_j R_j}}} \longleftarrow \boxed{\left(\frac{E_b}{N_0}\right)_j = \frac{W}{R_j} \bullet \frac{P_j}{I_t - P_j}}$$

$$\boxed{\begin{array}{lcl} R_j \uparrow & \Rightarrow & L_j \uparrow \\ \left(\frac{E_b}{N_0}\right)_j \uparrow, PLR \downarrow & \Rightarrow & L_j \uparrow \end{array}}$$

(E_b/N_0)$_j$	*: Energy per bit per noise spectral density for connection j*
P_j	*: Received power for flow j*
R_j	*: Guaranteed bitrate for flow j*
W	*: CDMA chiprate*
I_t	*: Total received power including thermal noise power*
L_j	*: Load factor for flow j*
η	*: Total load factor*

Homework

1. Review slides
   » use them to guide you through the recommended book
2. Read from *G. Miao* - *Fundamentals of Mobile Data Networks*
   » Chap. 4 - Scheduling
3. Answer questions at moodle