Priority Groups

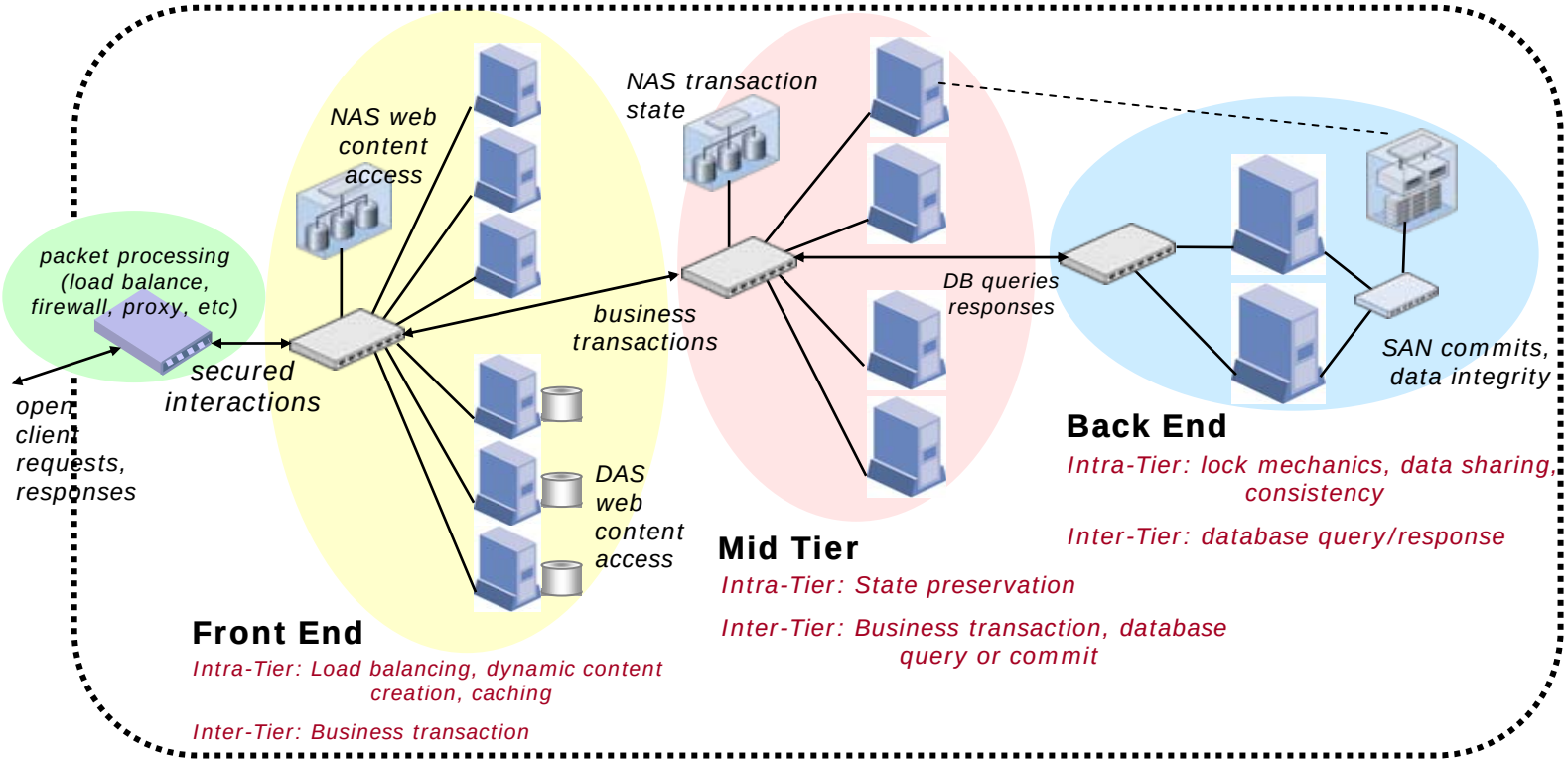
(Traffic Differentiation over converged link)

Manoj Wadekar (Intel)
Mike Ko (IBM)
Ravi Shenoy (Emulex)
Mukund Chavan (Emulex)

Agenda

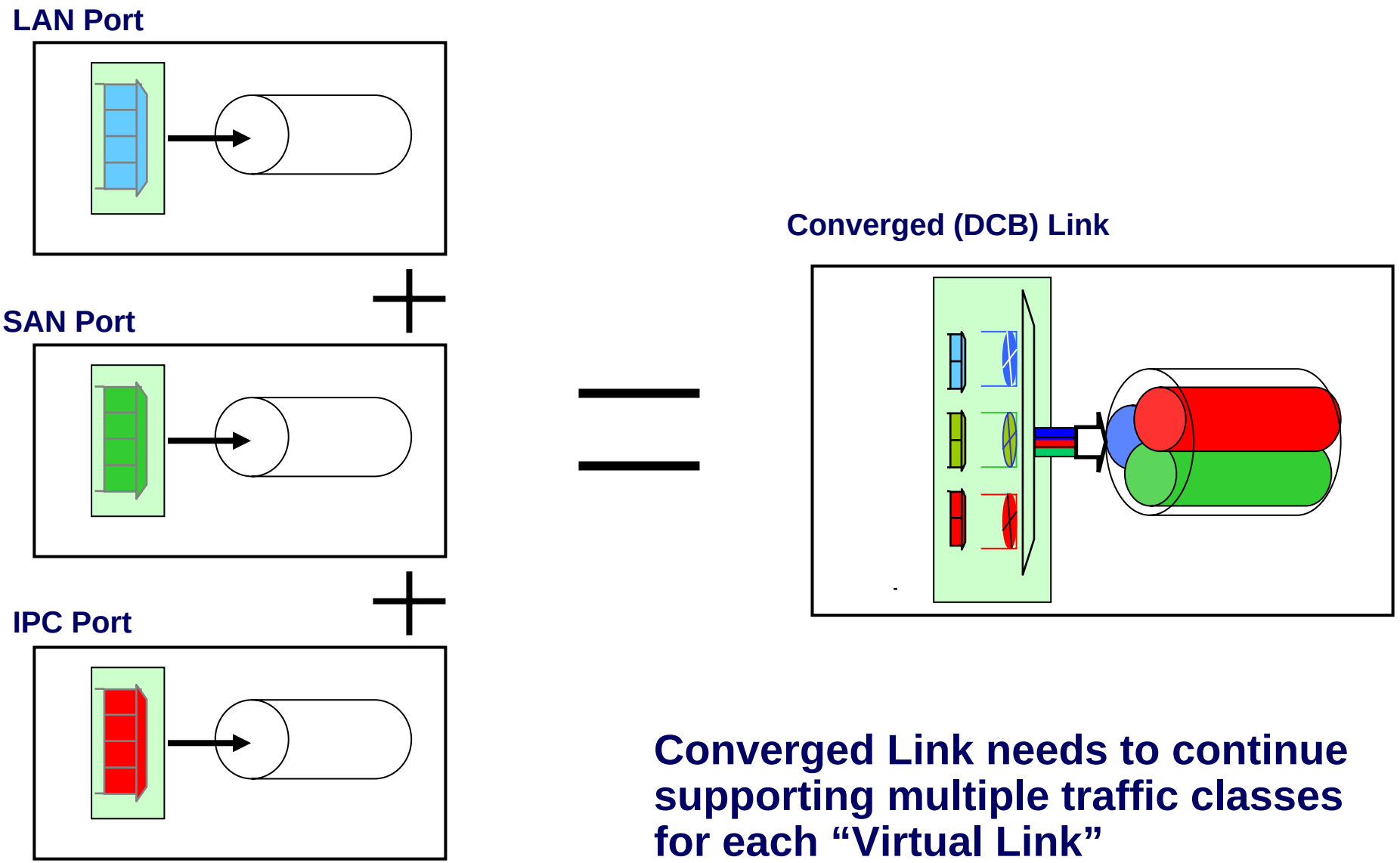
- **Usage Model**
- **Requirements – re-emphasized**
- **Configuration Tables**
- **Template config tables**
- **Summary**

Data Center Topology and Workloads



- **LAN:**
  - Legacy, bulk traffic: e.g. web access, email, file transfer
  - High priority, latency sensitive traffic – e.g. VoIP, Video-over-IP
  - Low priority, high BW traffic: e.g. back-up traffic
- **SAN:**
  - High BW, “no drop” traffic
  - Most of the traffic is between initiators and targets, not between servers
- **IPC:**
  - High BW data traffic
  - Low BW “latency sensitive” traffic
  - Lot of server-server traffic

DCB: Converged Link



Converged Link needs to continue supporting multiple traffic classes for each “Virtual Link”

Management view of converged link

- **DCB Cloud has multiple devices that support converged links**
  - Provide consistent management hooks
- **Configuration for BW assignment for each “Priority Group”**
  - Example: 40% LAN, 40% SAN, 20% IPC
- **Should allow multiple traffic classes within “Priority Group”**
  - Allows these traffic classes to share BW without hard configuration
  - Example: VoIP and Bulk traffic to share 40% LAN BW
- **Can not compromise low latency application due to convergence**
  - MUST allow strict, high priority scheduling of IPC (and equivalent) traffic
- **Should provide management infrastructure (MIBs)**
  - Defining scheduling algorithms is too restrictive and not necessary
  - Interoperability for management is important

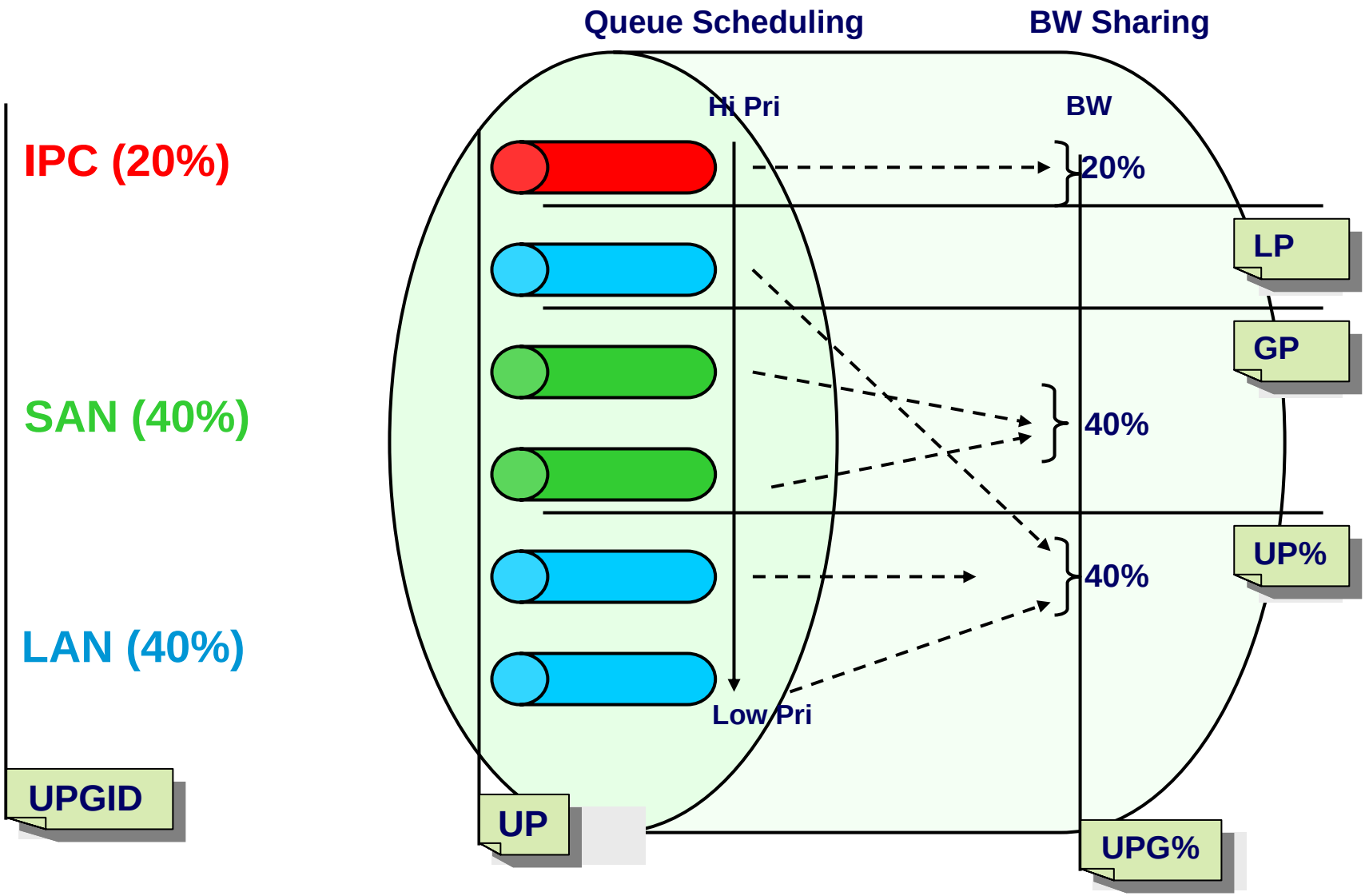
Goal and Non-goal

- **Goal for following slides is to kick-off discussion**
- **It is not intended to propose a solution for 802.1Qaz adoption**
- **No intention to propose scheduling algorithm**

Term Definition

- **UP: User Priority**
  - This is actual marking of traffic on the wire (802.1p bits)
- **User Priority Group (UPG) - UPGID**
  - E.g. LAN, SAN, IPC, Management etc.
- **UPG%**
  - % of Link Bandwidth allocated for a particular UPGID
- **UP%**
  - % of Group Bandwidth allocated for a particular UP within UPGID
- **LP (Link Priority)**
  - No BW check for this priority – follows strict priority scheduling
- **GP (Group Priority):**
  - If non-strict-priority scheduling is provided within a group, then this bit provides overriding strict priority behavior for given UP in the group

Traffic type, queues, scheduling, BW



Configuration Tables:

UP	UPGID	LP	GP±	UP% ±	Desc
0	2	False	True	-	LAN
1	2	False	True	-	LAN
2	1	False	True	-	SAN
3	1	False	True	-	SAN
4	2	False	True	-	LAN
5	2	False	True	-	LAN
6	NC	NC	NC	NC	NC
7	0	True	-	-	IPC

Table 1: UP-UPGID Table

UPGID	UPG%	DESCRIPTION
0	-	IPC
1	50	SAN
2	50	LAN
-		
-		
-		
-		

Table 2: UPG-BW Table

±: To be used if group uses non-strict-priority scheduling

Config Example: Strict Priority

UP	UPGID	LP	GP±	UP% ±	Desc
0	0	True	-	-	DEF
1	0	True	-	-	DEF
2	0	True	-	-	DEF
3	0	True	-	-	DEF
4	0	True	-	-	DEF
5	0	True	-	-	DEF
6	0	True	-	-	DEF
7	0	True	-	-	DEF

Table 1: UP-UPGID Table

UPGID	UPG%	DESCRIPTION
0	100	DEFAULT
-	-	-
-	-	-
-		
-		
-		
-		

Table 2: UPG-BW Table

Config Example: WRR/DWRR

UP	UPGID	LP	GP	UP%	Desc
0	0	False	False	12.5%	DEF
1	0	False	False	12.5%	DEF
2	0	False	False	12.5%	DEF
3	0	False	False	12.5%	DEF
4	0	False	False	12.5%	DEF
5	0	False	False	12.5%	DEF
6	0	False	False	12.5%	DEF
7	0	False	False	12.5%	DEF

Table 1: UP-UPGID Table

UPGID	UPG%	DESCRIPTION
0	100	DEFAULT
-	-	-
-	-	-
-		
-		
-		
-		

Table 2: UPG-BW Table

Config Example: WRR/DWRR with one Strict Priority Queue

UP	UPGID	LP	GP	UP%	Desc
0	0	False	False	14.2%	DEF
1	0	False	False	14.3%	DEF
2	0	False	False	14.3%	DEF
3	0	False	False	14.3%	DEF
4	0	False	False	14.3%	DEF
5	0	False	False	14.3%	DEF
6	0	False	False	14.3%	DEF
7	1	True	-	-	IPC

Table 1: UP-UPGID Table

UPGID	UPG%	DESCRIPTION
0	100	DEFAULT
1	-	IPC
-	-	-
-		
-		
-		
-		

Table 2: UPG-BW Table

Summary

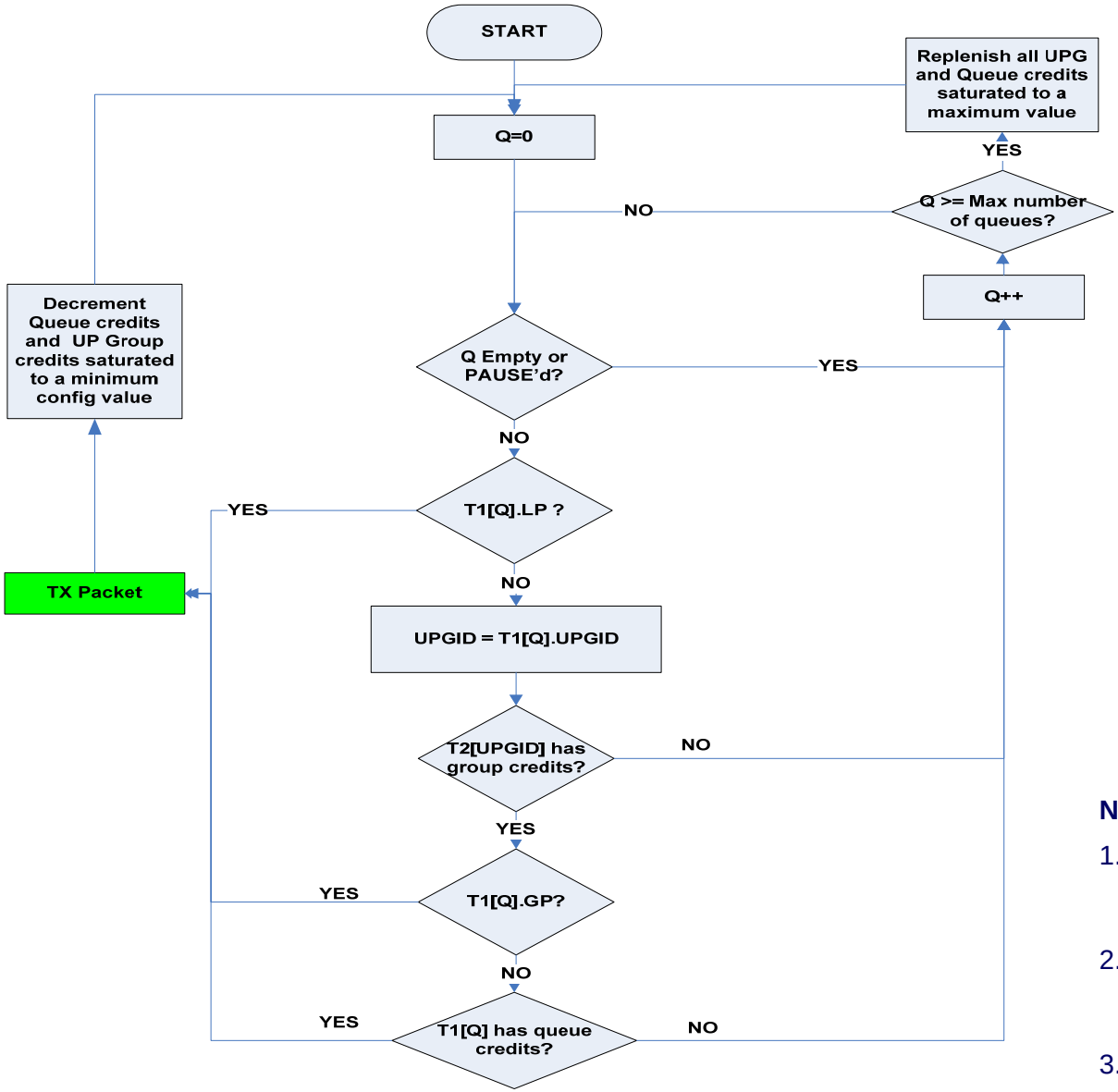
- **Allow BW configuration for Traffic Classes**
- **Consistent configuration mechanisms across devices**
- **Maintain low latency treatment of certain traffic classes**
- **Allow configuration of converged link to support BW sharing**
- **Maintain flexibility of implementation algorithms**

Backup

Device Configuration Mapping:

- **TC: Traffic Class**
  - This is specific to a device and maps into queues on egress ports
  - Could be less than number UP's on wire
  - Mapping is provided by MIB configuration
- **TCG:**
  - Group of traffic classes – derived from UPG
- **TCG%:**
  - % of Link Bandwidth allocated for TC group
- **TC%:**
  - % of Group Bandwidth allocated for particular traffic class
  - Multiple UP's may be concatenated in single TC
  - Mapping of Q% to TC% follows UP <-> TC mapping
- **TCLP:**
  - LSP mapping for TC
  - If multiple UP's are mapped to same TC, then behavior must be defined
- **TCGP:**
  - GSP mapping for TC
  - If multiple UP's are mapped to same TC, then behavior must be defined

Sample example , not proposal for standardization:



Notes:

1. Scheduling works on Traffic Class and hence config of UP needs to be mapped to TC
2. Assume 1:1 mapping of UP to TC and each TC is identified here with "Q"
3. T1: Table 1
4. T2: Table 2