Layer 2 power management proposal

Hugh Barrass (Cisco)

(also David Law - 3 Com)

Re-statement of July proposal

Significant changes:

Adoption of LLDP as the underlying protocol

- Relying on burst mode changes to overcome previous objections

Use of LLDP-MED as a starting point

- Logical step, given change to LLDP

Requirements of stateful management

(from barrass_2_0506.pdf)

State definitions

- Should be small # of states
- Power mode – consists of set of operating information
- States define how to set or change power mode

Robust state change mechanism

- Need to control state changes…
- … and be sure of partner's state
- Needs data transfer protocol
- Prefer defined and well-known protocol
- Compatibility or similarity to other standards (TR41) a plus

Should allow simplifications

- Trade cost vs optimization

Communication protocol

802.1AB LLDP

- Published 2005
- Currently under revision (will allow burst usage)
- Defined protocol frames
  - Very small impact on data b/w (1 per second regularly)
  - Not forwarded by bridges (needs definition in TPMR)
- Define new TLVs code (eventually add to Annex G)

Use a heart-beat communication mechanism…

- Periodically send state
- Mode change via request and acknowledge states
- Needs burst of messages

… could also include alarms for sudden changes

- Can communicate emergency situations (e.g. dying gasp)

How to use .1AB LLDP

- Normative references
  - Require support for LLDP (may be optional / mandatory)
  - Define objects for power mode
  - Define TLVs & codes for state communications
  - Eventually fold into 802.1AB Annex G (not necessarily during current revision)
- Also include informative
  - 802.3at Annex showing entire LLDP frames for mode change communications
- Other consideration
  - Whether to allow operation using alternate state communication?
  - Extend TIA TLV or 802.3 TLV – must consider co-existence
  - This presentation assumes all new TLV's use 802.3 org. code
- States & communication mechanism first priority
  - Focus on states first – collect power mode objects later

Power mode objects (PSE & PD) for example

Some suggestions for discussion…

- Not necessary for state & communication definition

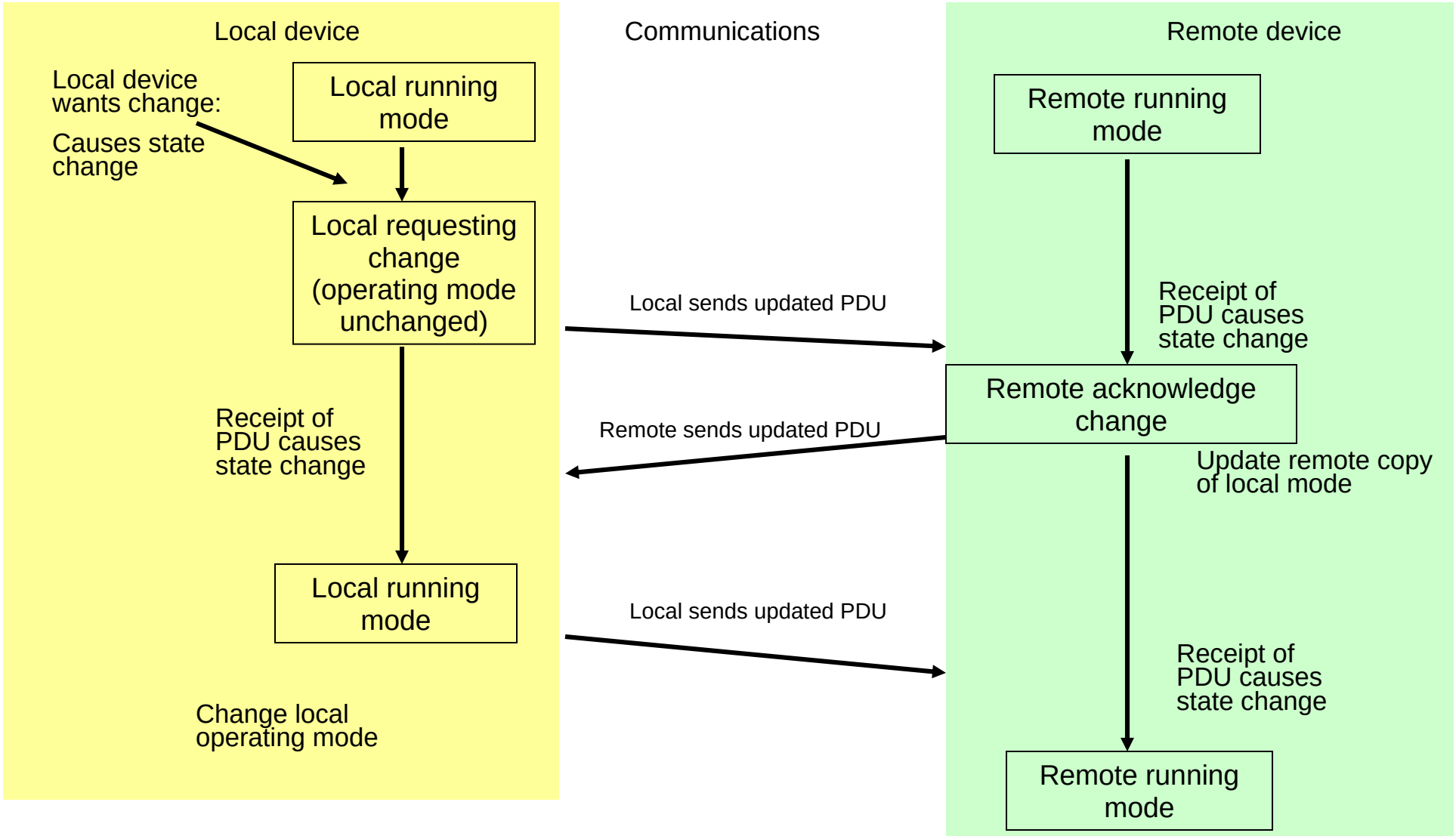
Power mode (actual and requested)

- actualPeakPower; actualAveragePower; remainingPowerMargin; requestedPeakPower; requestedAveragePower
- (average could be defined as 10 second moving average)
- Other objects TBD; e.g. support for statistical oversubscription

Plus, of course state definitions PSE & PD

- State: running; requestingNewMode; aknowledgeChange; non-acknowledge – maybe some others for startups & exceptions

General state change procedure



Detailed state behavior (1)

Periodically send state summary PDU (constantly)

- i.e. once per 30 second send PDU & process received

If local in running state and remote changes to requesting state

- Observe remote requestedPower objects
- Change to acknowledge or non-acknowledge state
- (depending on acceptance of change)
- If acknowledge, change local copy of remote actualPower objects
- Send PDU reflecting new settings

When remote state changes to running

- Change to running state
- Send PDU reflecting new settings

Detailed state behavior (2)

If local in running state and local device wishes to change

- (requires most recent remote PDU must be in running state)
- Set local requestedPower objects
- Change to requesting state
- Send PDU reflecting new settings

If remote changes to acknowledge state

- Change operating power mode; update local actualPower objects; change to running state
- Send PDU reflecting new settings

Else if remote state changes to non-acknowledge

- Do not change operating power mode; change to running state
- Send PDU reflecting new settings

Detailed state behavior (3)

Collision event

- Local in requesting state, remote changes to requesting state
- Do not change operating power mode; change to running state
- PD waits before repeating request (PSE does not need to wait)

Initial state

- After power up, use L1 classification as first actualPower mode (both local and remote)

Loss of communication

- NB – L1 classification has precedence (for connect & disconnect)
- If PDU received after TBD time assume dead partner
- Procedure TBD if partner dead – to define explicitly for PSE & PD
- Also need behavior for unexpected state change

Persist or vacillate

State definitions require that each request must be acknowledged or denied before returning to running state

- The requestor must not de-assert request until ack/non-ack
- The partner must respond to request as soon as it is seen

The requestor may persist or vacillate after non-ack

- To persist, simply re-assert request – after TBD delay
- Or decide not to persist (withdraw request)

NB some offers cannot be refused!

- Regardless of acknowledge, PSE might withdraw power if necessary

More stuff…

Do we need a “deep sleep” mode?

- PD does not communicate, maybe link down
- PSE allocates enough power to restart

Graceful power withdrawal – allows clean shutdown

- PSE request change to 0 power
- Allows PD to indicate controlled power-down

Minimal PD behavior – no state change supported

- => always stick in initial state
- (Equivalent to .3af for lower power)
- Simplest management = request 1 operating mode & stick there
- (no possibility of PSE requested change)

PoE mode LLDP-MED frame

PDU frame – detailed contents (suggestion…)

Bytes	Content	Value	Description
12	MAC DA / SA	01-80-C2-00-00-OE / ?	LLDP DA / sender address
2	Ethertype	88CC	LLDP type
9	Mandatory TLV		Chassis ID TLV
9	Mandatory TLV		Port ID TLV
4	Mandatory TLV		TTL TLV
2	TLV type/len	127 / 7	State code
4	OUI / subtype	00-12-BB / 04	TIA OUI / PoE subtype
1	Type / source / PRI		Power type, source and priority
2	Power		Power level 0 – 102.3 Watts
19	PAD / FCS		

New TLVs

In addition to LLDP-MED TLV – send extra info in new TLVs

- LLDP-MED power & pri = “actual” (peak) power & priority

Use 802.3 OUI

- New subtypes (to be added to Annex G)

New TLV includes: Requested power mode, power state

- State: running; requestingNewMode; aknowledgeChange; non-acknowledge
- Power: remainingPowerMargin; requestedPeakPower;

Additional TLV for other info:

- actualAveragePower; requestedAveragePower
- (average could be defined as 10 second moving average)
- Other objects TBD; e.g. support for statistical oversubscription

Projected – New TLV

Power state TLV – detailed contents (suggestion…)

Bytes	Content	Value	Description
1	TLV type	127	Organizationally dependent
1	length	10	Length
3	OUI	00-12-0F	IEEE 802.3 OUI
1	subtype	05	New subtype – POE state transfer
1	power state		Enumerated: running; requestingNewMode; aknowledgeChange; non-acknowledge
			(could include “urgency indication” for dying gasp)
2	requested power		Requested peak power (0 – 102.3 W)
1	requested priority		Requested power priority
2	remaining power		Remaining power margin (PSE only : 0 – 102.3 W)

Projected – Additional TLV

Optional power information – detailed contents (suggestion…)

Bytes	Content	Value	Description
1	TLV type	127	Organizationally dependent
1	length	7	Length
3	OUI	00-12-0F	IEEE 802.3 OUI
1	subtype	06	New subtype – POE additional information
2	Actual average power		(0 – 102.3 W) – averaged over 10 sec
2	Requested average power		(0 – 102.3 W) – subject to state change
			more could be inserted here …

Next steps

- Turn this into a complete normative definition
  - Not feasible in PowerPoint
  - Clause changes – locations & text
- Write informative descriptions of frames
  - For each key frame type
- Make list & definitions for power mode objects
  - Primary work for power experts – not dependent on mechanism
  - Possible starting list…
  - actualPeakPower; actualPowerPriority; (in LLDP-MED)
  - requestedPeakPower; requestedPowerPriority; remainingPowerMargin;
  - actualAveragePower; requestedAveragePower
- Address TBDs
  - Loss of communication (PSE & PD) behavior; deep sleep mode; simplified PD behavior; etc.

Questions…

… or comments

(partial) list of supporters

The following people have indicated that they support the adoption of this slideset as the baseline for the L2 management mechanism.

Wael Diab (Broadcom)

Hugh Barrass (Cisco)

Baseline adoption motion

Move that the Task Force adopt barrass_1_0107.pdf as the baseline for the L2 management mechanism for 802.3at (not including the TLV definitions)

P: Hugh Barrass

S:

Y: nn N: n A: n

(802.3) Y: nn N: n A: n