draft-ietf-bess-ebgp-dmz

S. Litkowski
S. Mohanty
A. Vayner
A. Gattani
A. Kini
J. Tantsura
R. Das

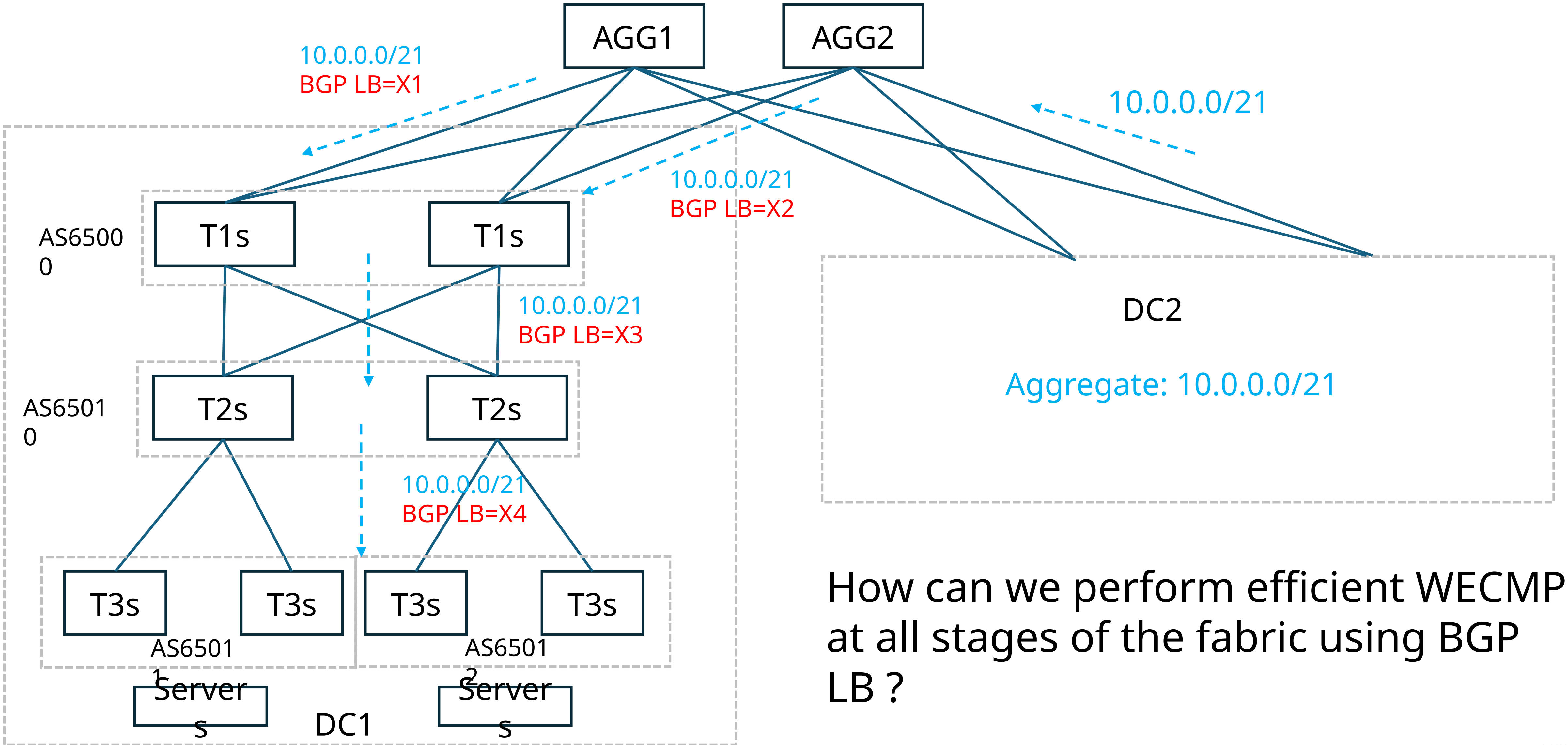
Reminder

- Draft focus:
  - use cases of BGP link bandwidth community
  - Local mechanisms to tweak WECMP and BGP link bandwidth community settings (e.g.: cumulation)

Main updates (details after…)

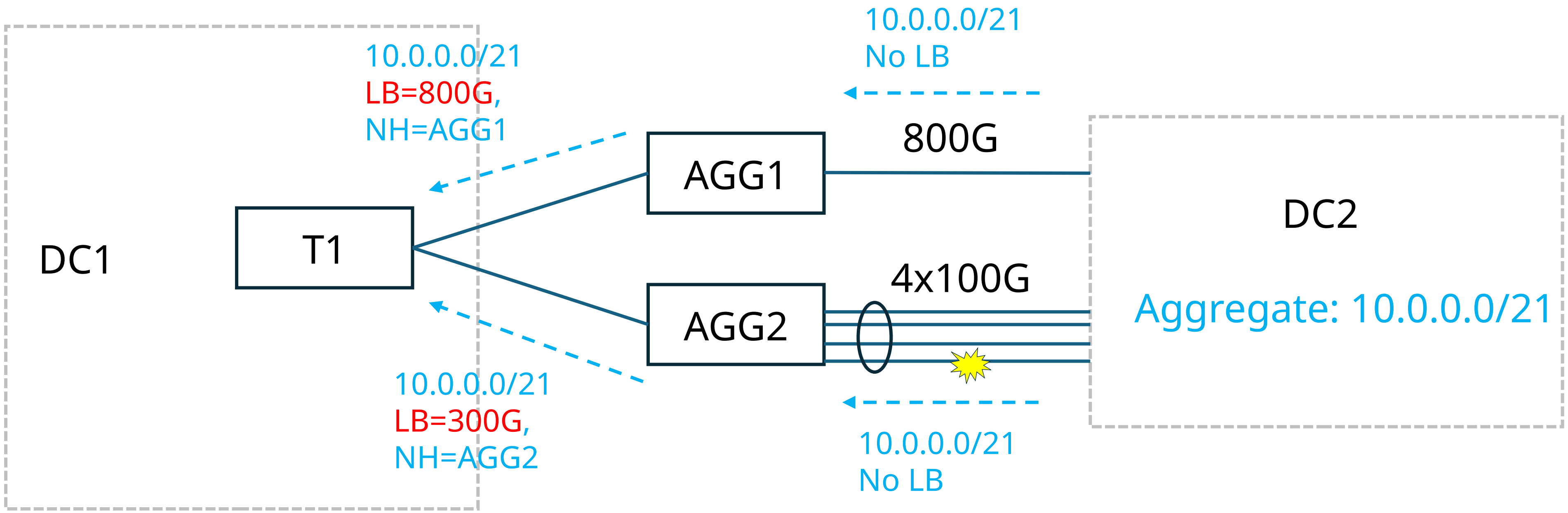
- Clean up inaccurate text related to BGP link bandwidth draft
- Clear separation between use cases and behaviors
- Behaviors defined: (reflecting what is deployed in field)
  - Link bandwidth discovery
  - Contributing link bandwidth usage for WECMP
  - Cumulation
- Proposal to make it back to standard track

Use case example: eBGP-based DC



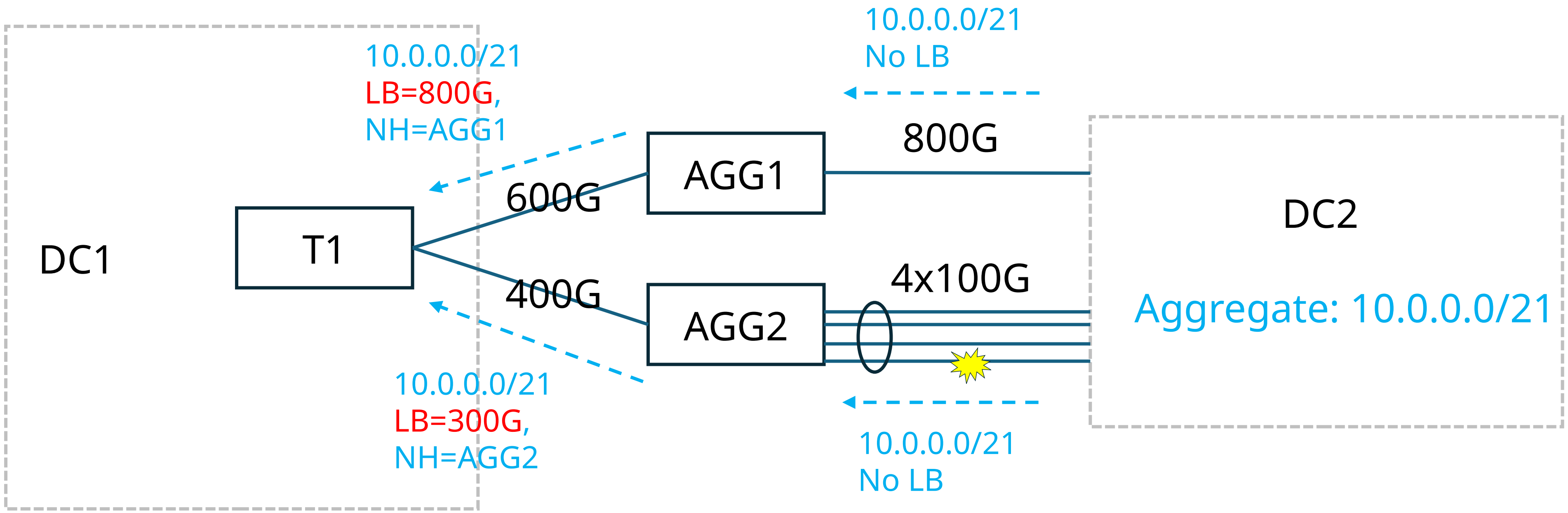
How can we perform efficient WECMP at all stages of the fabric using BGP LB ?

Initialization of the BGP link bandwidth



- Bandwidth to be used in BGP LB can be learned from interface bandwidth on which the eBGP peer is established
- If interface bandwidth changes, BGP LB value is updated (bandwidth tracking)
- Enabled by configuration

WECMP considerations: contributing bandwidth



T1 has two BGP paths for 10.0.0.0/21 with resp BGP LB of 800G (AGG1) and 300G (AGG2).
T1 can reach AGG1 over a link of 600G and AGG2 over a link of 400G.

Using ratios based on BGP LB values will create a congestion.

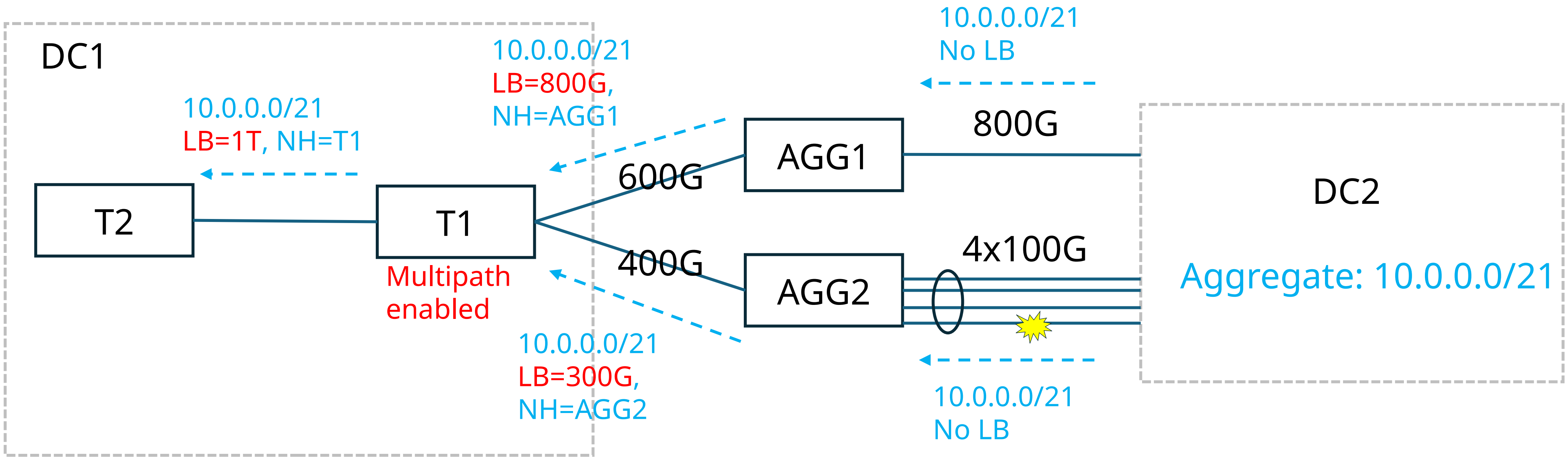
T1 may consider (for example) the minimum between the BGP LB value and the local link BW to compute WECMP

WECMP considerations: contributing bandwidth

- Definitions:
  - Remote bandwidth = bandwidth value learned from BGP LB
  - Local link bandwidth = bandwidth value of the link of the connected peer

  - Contributing bandwidth = bandwidth value used to compute WECMP weights
    - Value is set from:
      - Remote bandwidth value
      - Local link bandwidth value
      - Function using both remote and link bandwidth (e.g.: min(remote, local))
    - Controlled by local configuration

Cumulation



- As T1 performs WECMP and considers a bandwidth of 600G to AGG1 and 400G to AGG2
- In total, T1 has 1T of BW to reach DC2
- When advertising DC2 aggregate down to T2, T1 is cumulating the BW available from the multipath
- Cumulation is enabled by configuration (may be selective)
- Cumulation may be done based on:
  - Contributing bandwidth OR local link bandwidth OR remote bandwidth

Next steps

- Feedbacks on the new version are welcome
- Are we good with standard track ?
- Need to discuss move of the draft to IDR