NVO3: Network Virtualization Problem Statement

Thomas Narten
narten@us.ibm.com

Interim WG Meeting - Westford, MA
Sept 20, 2012

Where We Are At

- Had both problem statement and VPN4DC draft in Vancouver
- After Vancouver, parties worked together to produce single problem statement, posted Aug 10
- Successful call for WG adoption, WG ID posted Sept 5
- Eric Gray/Thomas Narten are document editors
- Several additional reviews that need to be folded in, edits are mostly editorial, will post after meeting
- Plus whatever edits come out of this meeting...

Discussion Topics (one per chart)

Data Plane

- Comment from reviewer: no text about data plane in problem statement
- Was some text early on, got pulled out when framework document came along
- There seems little need/desire to develop another encapsulation
- Almost any encap will do, so long as it has an acceptable Context ID/VN-ID
- Is it good enough to leave data plane discussion to the data plane requirements document?

Trombone Routing

- What is the definition of Trombone Routing in the context of overlays?
- By definition, intra-VN forwarding is forwarded directly
  - Or, is there weirdness with multi-subnet VNs?
- For inter-VN routing, one view:
  - Policy (by default) disallows communication
  - Policy may require traffic traverse an existing device (e.g., certified firewall)
  - Should NVO3 optimize the case where inter-VN communication is allowed?
- Process question: is this a “problem” that exists in today's networks that overlays can somehow solve?
  - Or is avoidance a solution requirement?

Ingress/Egress Path Optimization

- Goal: optimize paths through gateways to/from VN
- External device should use ingress gateway to VN that is “close” to VM
  - Multiple gateways may exist on same VN
  - VM may move over time (optimal ingress may change)
  - Overlay may have large breadth
    - 2 different gateways may be far apart physically
    - One may be much closer to target VM
- Goal: VM should use default router/egress gateway that is “close”, even after VM moves
- Complication: presence of middleboxes may pin traffic to a gateway for existing TCP flows
- Process question: Is this a “problem” or solution requirement?
- Note: root problem is IP; IP doesn't give visibility into a subnet – all nodes are “one hop” away

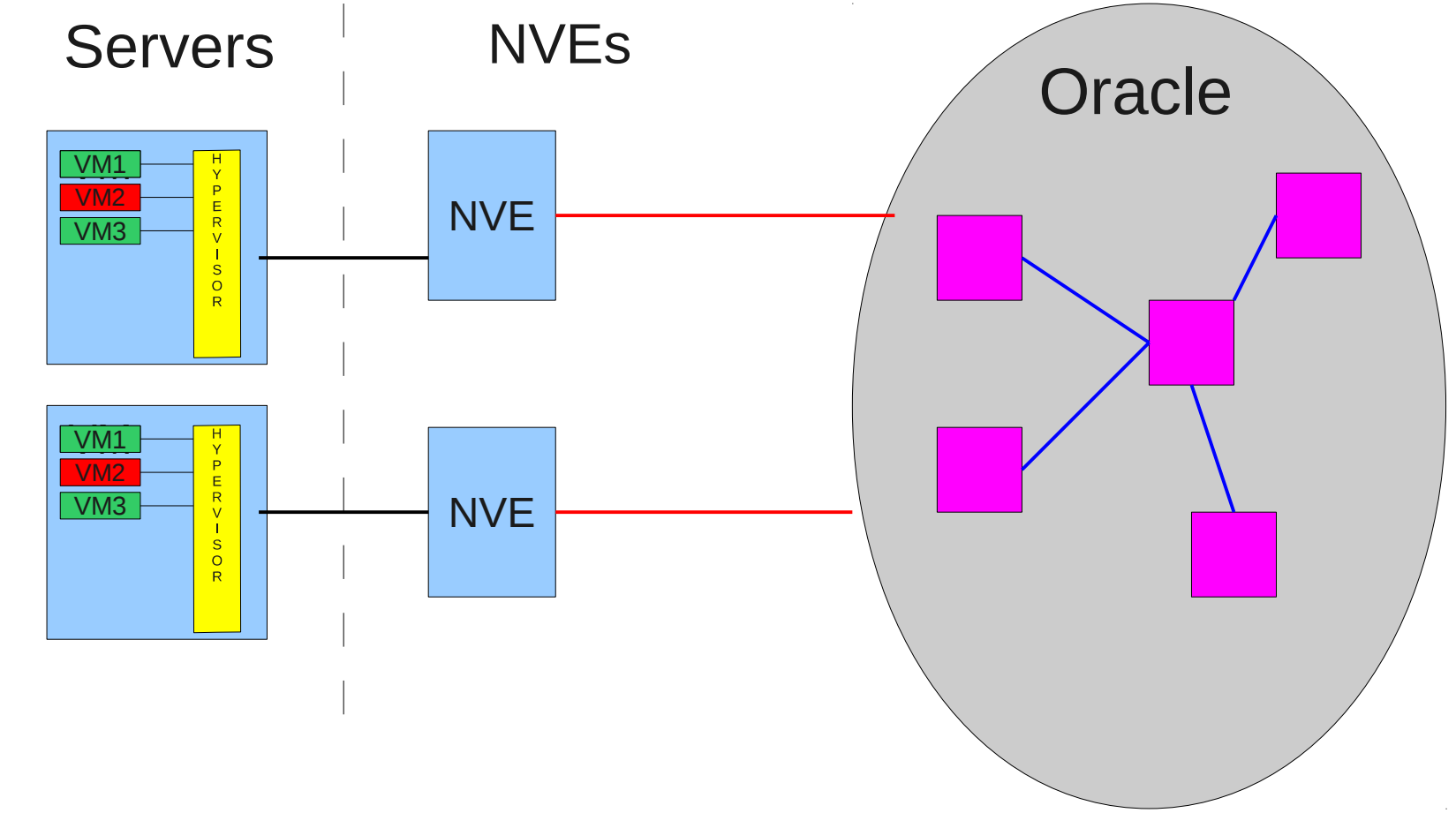
Discussion of L2 “Problems”

- Review by Janos Farkas highlighted imprecision and inaccuracies of text related to L2 “limitations”
  - e.g., not accurate to say VLANs limited to 4096
- Not sure we should remove everything about current L2 issues (but also need to be accurate!)
- One issue is difference between what is deployed and what has been developed but not (yet) widely deployed.
- L2 limitations are being felt in deployments now, and L3 overlays are one possible direction going forward
  - Other directions are possible too, but doesn't mean they are a sure thing either

NVO3 Work Areas (per draft)

- In discussions during and since Vancouver, it's been useful to clarify the potential control plane work areas
- Oracle Itself
  - Oracle - “a person or thing regarded as an infallible authority on something” [Oxford Dictionary]
  - In NVO3, NVE's can query the oracle to get whatever information they need to deliver traffic to remote VMs (e.g., inner to outer address mappings)
- NVE - Oracle interaction
  - The control protocol used between the NVE and Oracle
- Server - NVE interaction (in case where NVE is not co-located with server)

Three Potential NVO3 Work Areas



Oracle

- The “oracle” has full knowledge of all mappings and maintains/distributes such knowledge to NVEs
- Could be centralized/distributed/etc. - no presumption of how it is implemented, e.g.:
  - Based on an existing routing protocol (e.g., BGP, IS-IS, etc.), or
  - Implemented as a directory service
    - Existing VM orchestration systems already maintain centralized information about all VMs, IP/MAC addresses, current location, etc. - they are a logical place to implement an oracle
  - Something else?

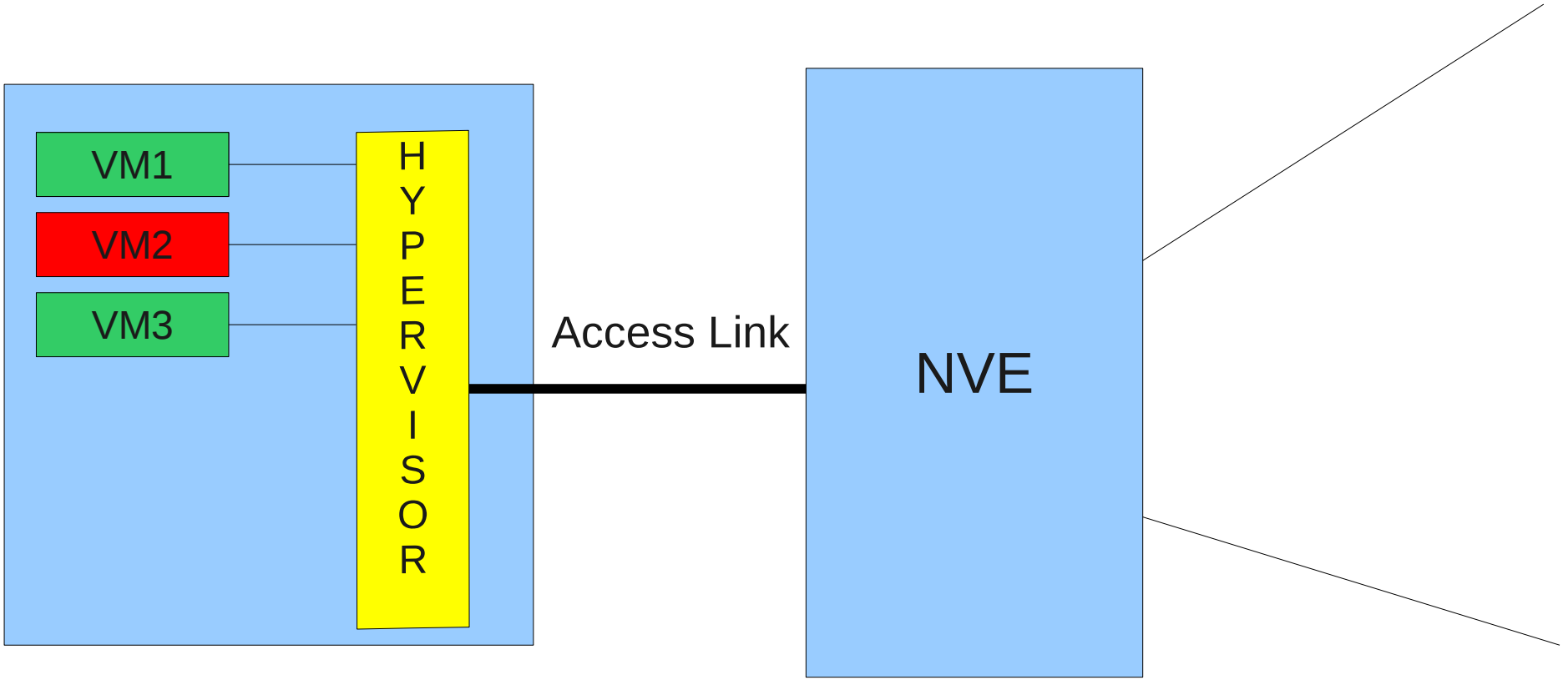
NVE-Oracle Interaction

- NVEs may pull information from oracle
  - E.g., NVE needs mapping for destination VM
- NVEs may push information to the oracle
  - E.g., A VM is attaching to (or detaching from) this NVE
- Oracle may want to push information to an NVE
  - The mapping for VMx has changed (VM has moved)
- Architectural choice:
  - NVE can just be part of the oracle, implementing same protocol (or subset) as oracle
    - But would tie the NVE implementation to oracle
  - Develop an oracle-agnostic, general-purpose NVE-oracle control plane
    - Allow NVE and oracle to evolve independently

Server-NVE Interaction

- When NVE is part of hypervisor all interactions are internal (no protocol needed)
- When hypervisor and NVE are on different devices separated by an access network
- Consider simple case (L2 Ethernet link)
  - Server/NVE will need to negotiate/agree on what VLAN corresponds to VM
  - NVE will need to be able to map VM/VLAN to VNI
  - Server needs to be able to inform NVE when VM attaches/detaches from VN
  - NVE may need to authenticate above operations

Server-NVE Interaction



Server-NVE Interaction

- How complex of an access link should we support?
- Always assume L2? (If IP, we'd need an IP encap that identifies the VN)
- VSI Discovery Protocol (VDP) is an existing IEEE protocol that may be leveraged for this purpose.
- Many proprietary protocols in this space already (between switch ports and NICs)

Questions/Comments

Acknowledgments

- Document co-authors: David Black, Dinesh Dutt, Luyuan Fang, Eric Gray, Larry Kreeger, Maria Napierala, Thomas Narten, Murari Sridharan
- Many helpful comments (on and off list)