Identifier Locator Addressing

IETF95

Tom Herbert <therbert@fb.com>

Drafts

- draft-herbert-nvo3-ila
- draft-herbert-ila-messages
- draft-lapukhov-ila-deployment
- draft-lapukhov-bgp-ila-afi

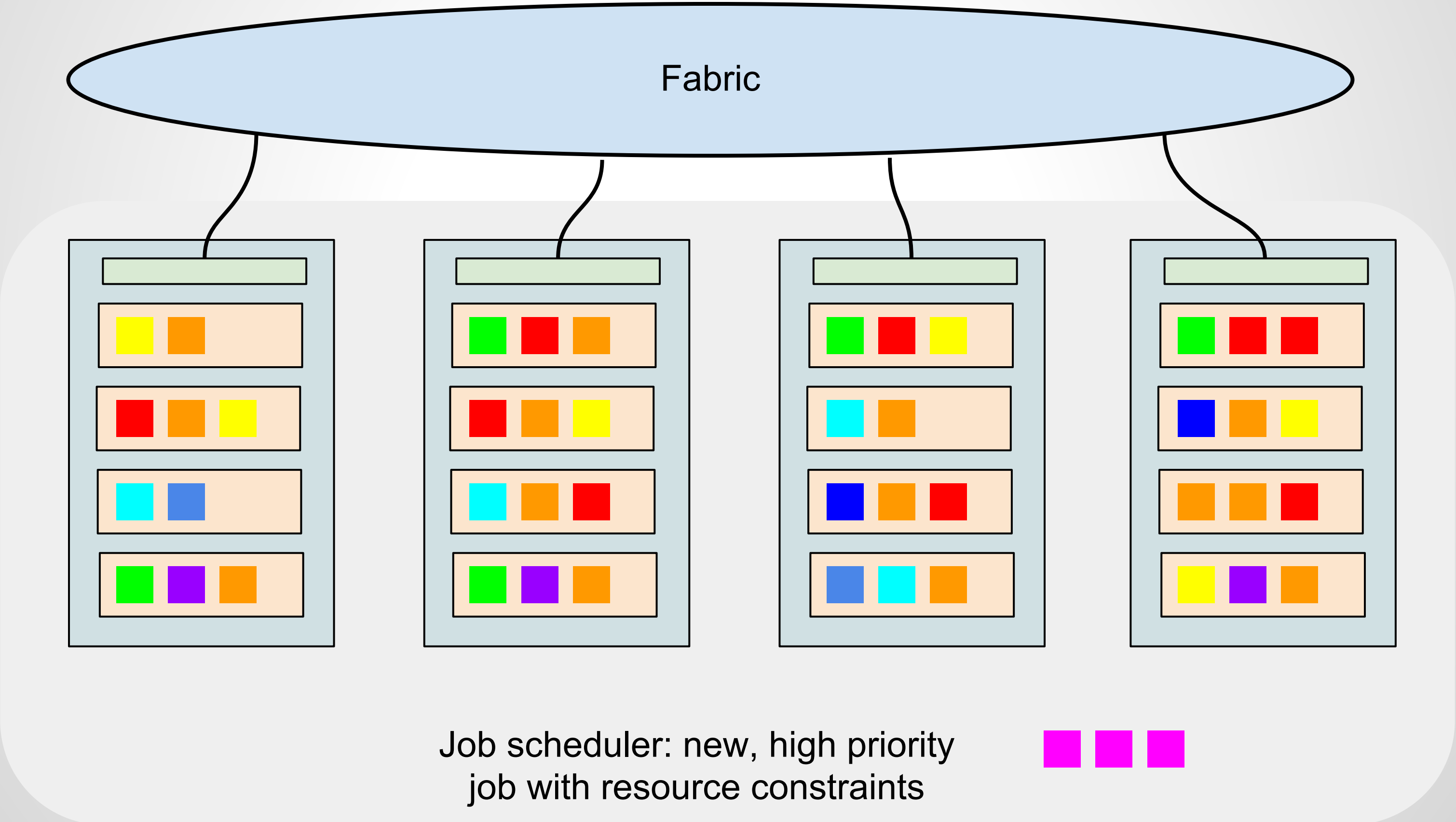
Motivation

- Object virtualization
  - Fine grained addressing of arbitrary objects
  - Support object migration between physical hosts
  - Scale to 10s or even 100s billion objects in DC
- Example
  - Virtualize tasks (containers)
  - Connectivity for VMs (external to VN)
  - IP Mobility (5gangip maybe)

Example: Task virtualization

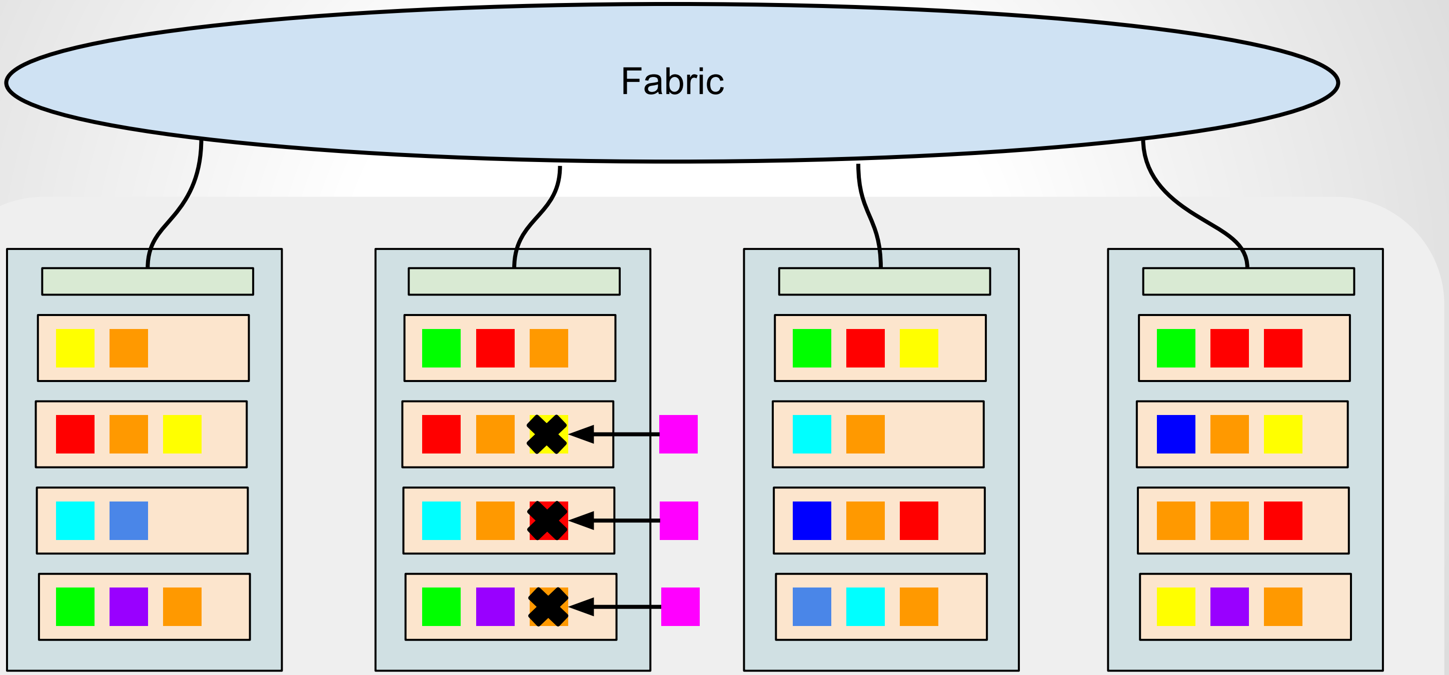
Capability that every task in the data center can be seamlessly live migrated per discretion of a job scheduler.

Scheduling dilemma



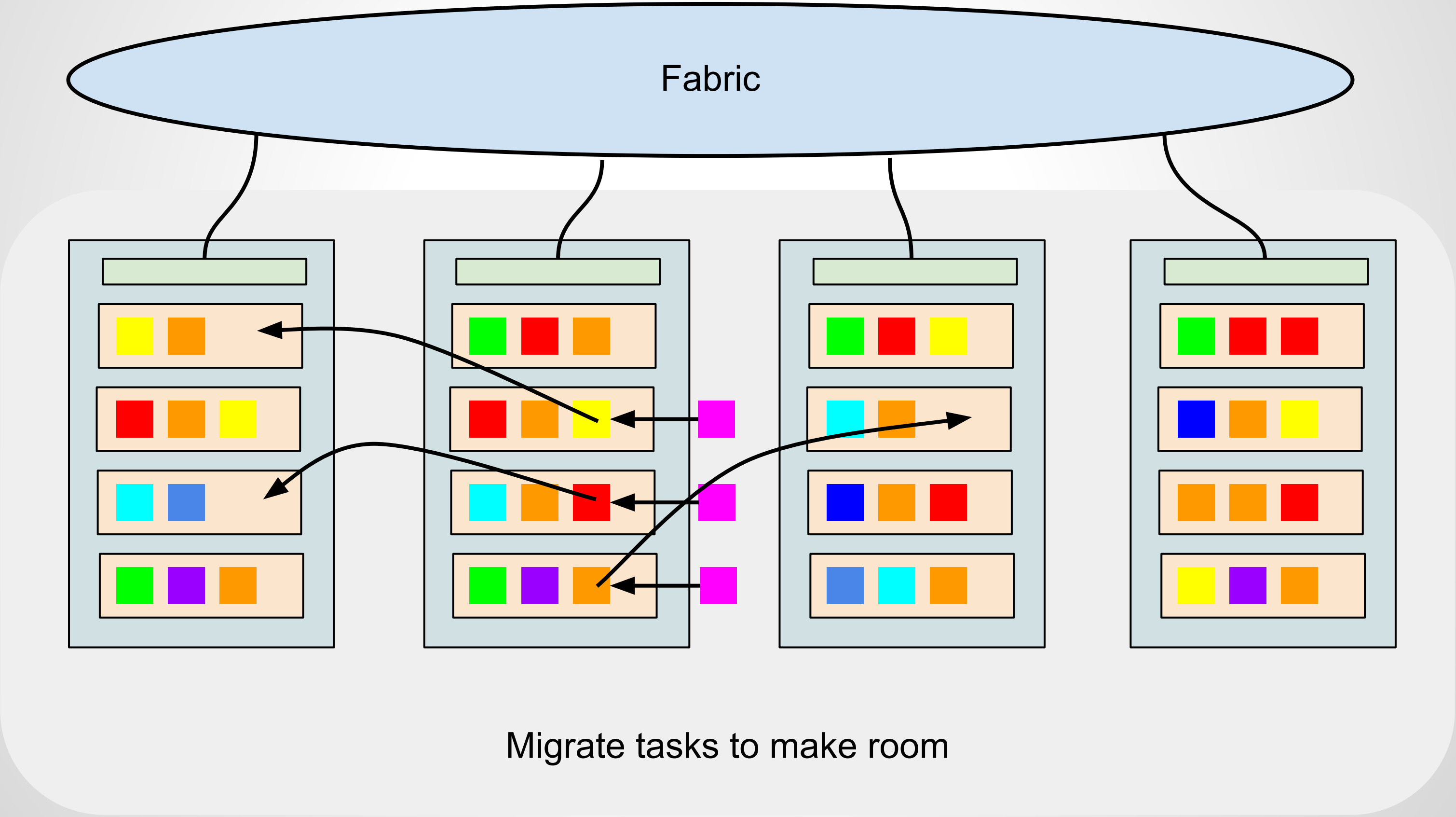
Job scheduler: new, high priority job with resource constraints

Unpleasant solution today



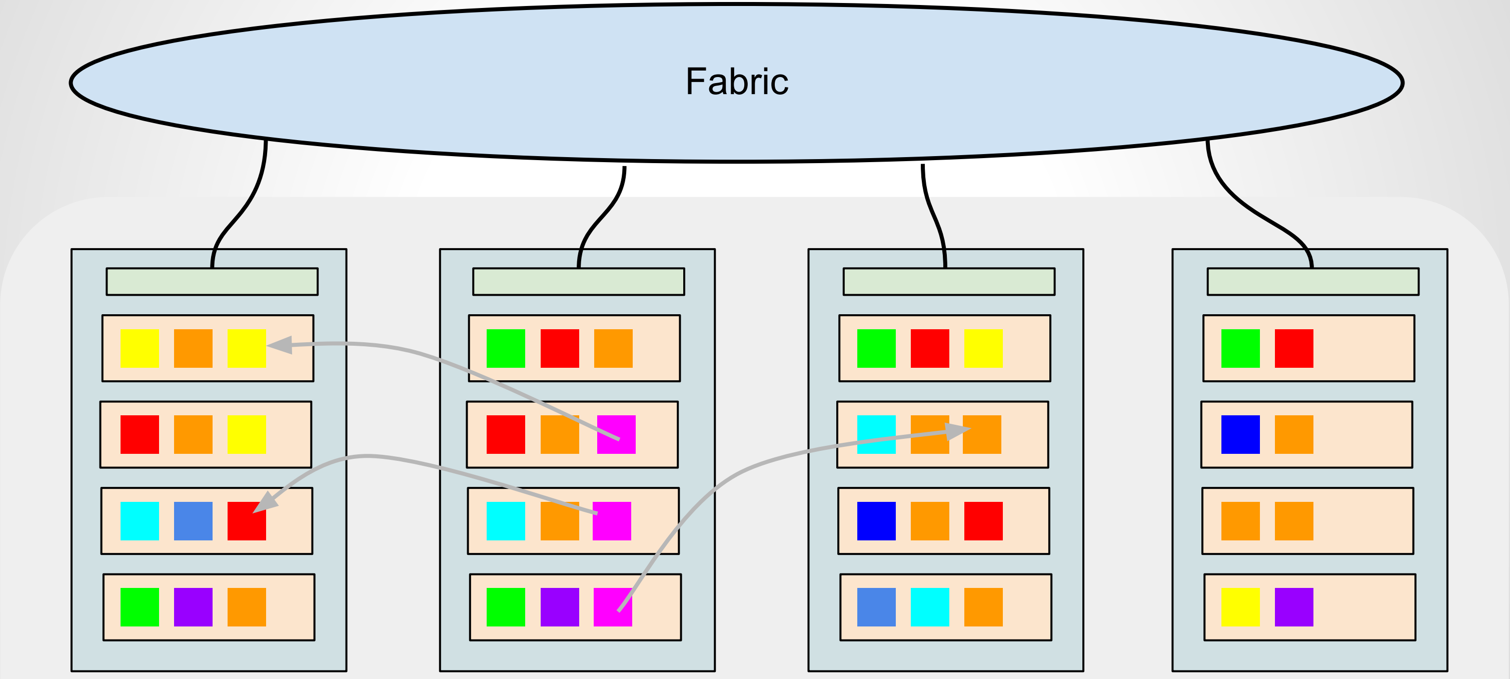
Kill existing tasks to make room

Task migration solution



Migrate tasks to make room

After migration



No tasks needed to be killed

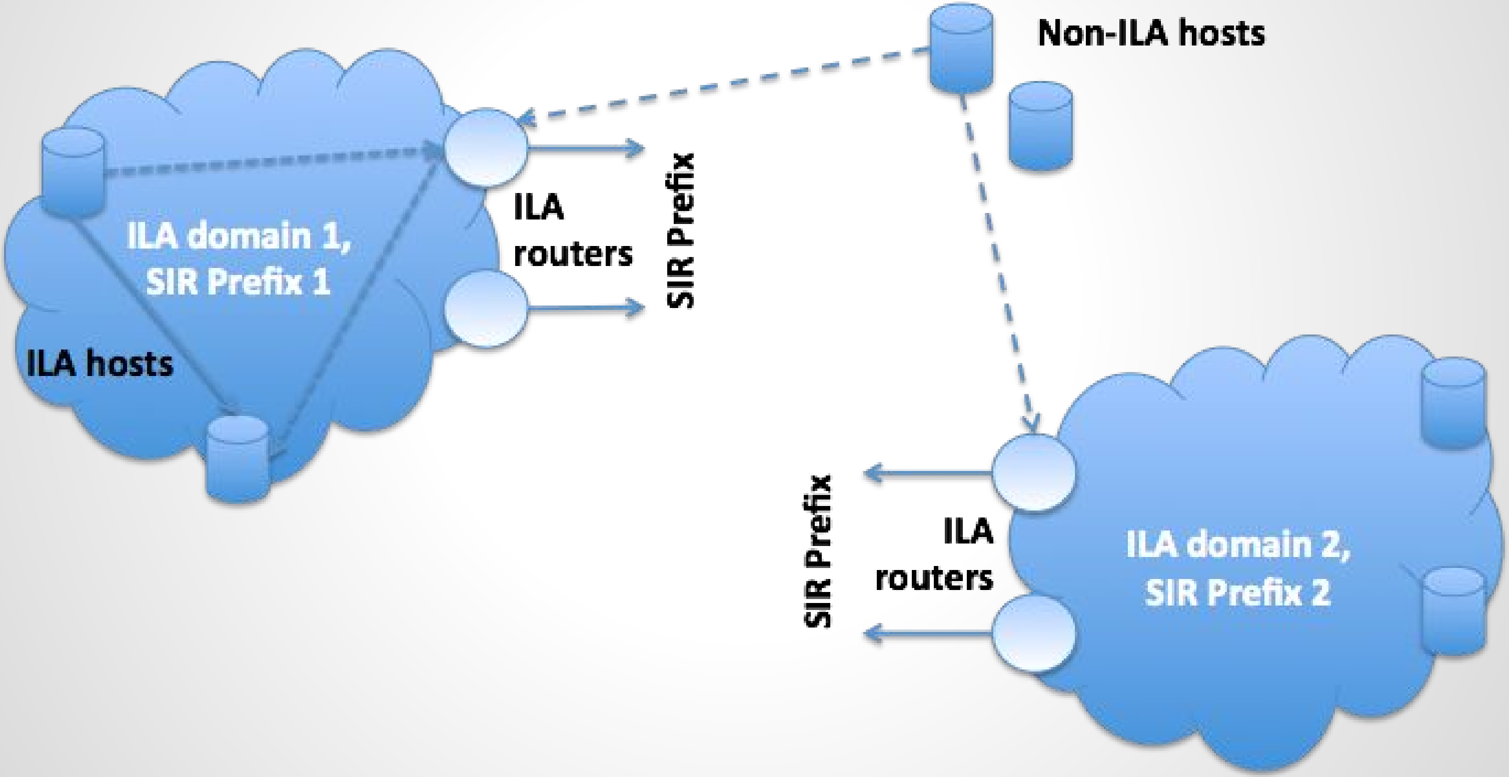
Requirements/assumptions

- Be **transparent** to apps, users, & network
- **Zero** performance impact when not migrating
- No on-the-wire overhead (i.e. no encapsulation)
- Does **not** adversely impact security or control
- No overlay networks, no vswitch needed
- ECMP and NIC offloads continue to work
- Most objects will probably never be migrated

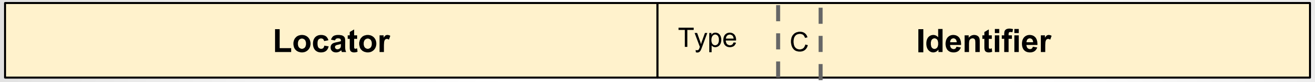
ILA Solution

- Split IPv6 address into identifier (who) and locator (where) ala ILNP
- Each object gets its own unique identifier
- Mapping identifiers to locators
- If object migrates between hosts, its locator changes but its identifier does not
- When not migrating, data path is essentially same as before

ILA topology



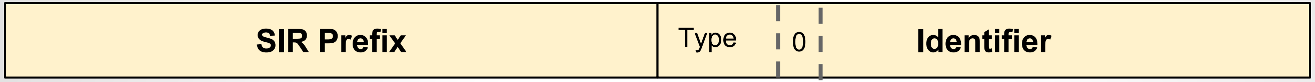
Address split



Locator | Type | C | Identifier

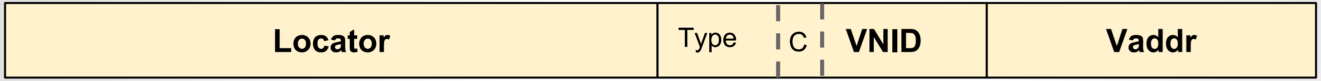
- Locator
  - 64 bits identifier of physical hosts
  - Routable
  - Not used as connection endpoint
- Identifier
  - 64 bit logical endpoint address of virtual node
  - Not routable
  - Used as connection endpoint
  - Typed to allow different modes

User Visible Addresses



- Standard Interface Representation (SIR)
    - A “virtual” address in ILA
    - Common SIR prefix in locator part of address
    - Applications, conn. endpoints use SIR address
- To actually route to destination SIR prefix is translated to locator per mapping table
- ILA translation assumed symmetric, both sides see same SIR addresses for an object

Network virtualization use cases



Locator	Type	C	VNID	Vaddr

- Embed VNID in ILA address
  - Potentially eliminate encapsulation for NVO3
  - No place to put security to authenticate VNID, so intra-VN use might be limited
- Allows VM to common DC service, or Internet w/o stateful NAT or encapsulation
- Allow two VMs to communicate under policy w/o NAT

Details

- Need to map identifiers to locators
  - Same problem of mapping Vaddr to Paddr in NV
  - Use NVO3 control plane to distribute mappings
- Translation can occur at end hosts or in network (since ILA only operates on L3)
  - ILA routers provide network translation service
  - “Redirects” can be sent by ILA routers to inform ILA capable hosts of locator so they can send directly

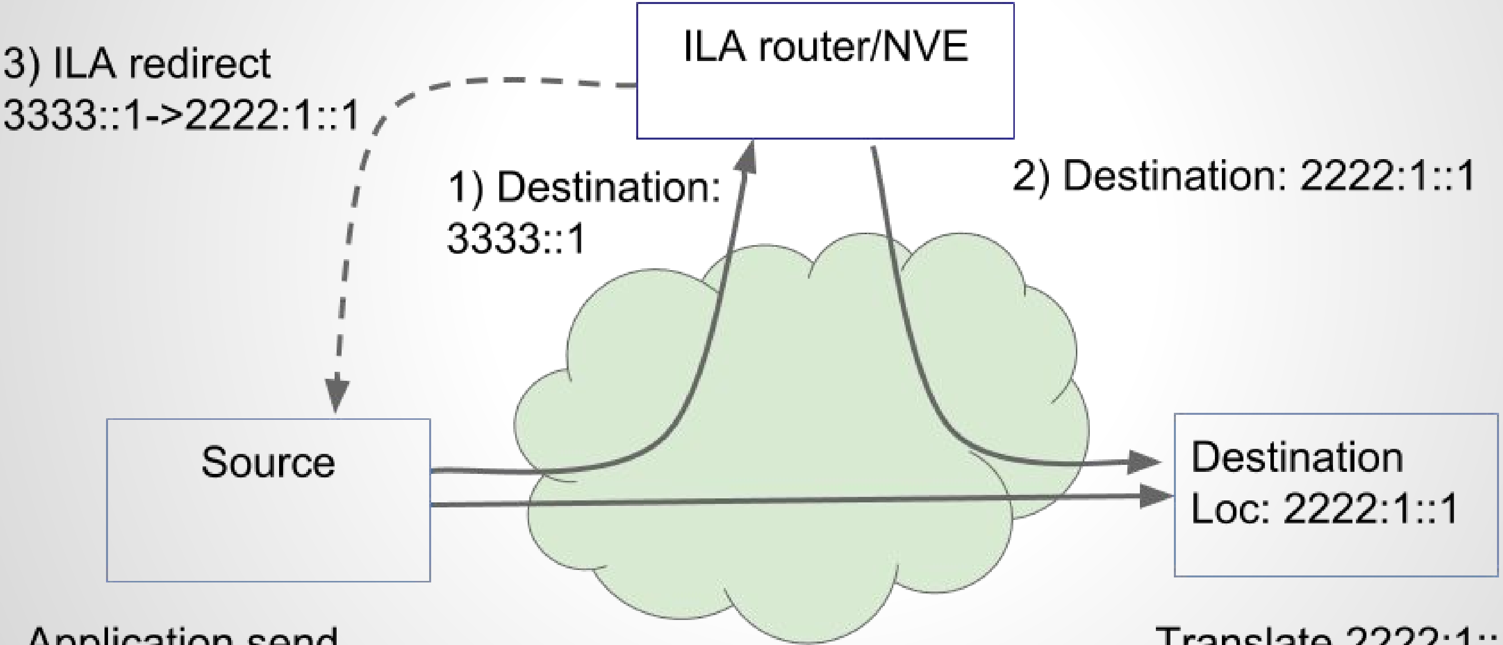
Identifier properties (60 bits)

- Uniqueness
- Not predicable
  - Given one know identifier, should not be able to predict what the next one created would be
- Example decentralized scheme (~RFC4122)
  - 24 bit host ID (each host autonomously creates IDs)
  - 36 bits obfuscated timestamp
  - Gives ~22 yrs. worth of identifiers before wraparound at 100 IDs created per second, per host

ILA routers

- ILA routers are assigned anycast SIR
- They translate SIR to locators for forwarding
- Map identifiers to locators, participate in a control plane to get this info
- “Redirects” are used to inform ILA capable hosts of ID->Loc mapping so they can perform translation directly

Communications flow



ILA router/NVE

1) Destination: 3333::1

2) Destination: 2222:1::1

3) ILA redirect 3333::1->2222:1::1

Source

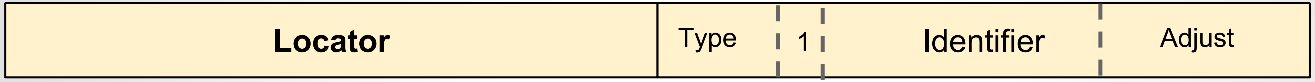
Destination
Loc: 2222:1::1

Application send to 3333::1 (e.g. from DNS)

4) Send directly to destination 2222:1::1

Translate 2222:1::1 to 3333::1, application receives from 3333::1

Checksum neutral translation



Locator	Type	1	Identifier	Adjust

- Like in RFC6296
- Format
  - C bit is set
  - Low order 16 bits of identifier
- On TX
  - Calculate adjustment based on 1s complement difference between old and new locator
  - Set C bit and Adjust field
- On RX do the reverse operation

Control plane

- Mapping dissemination among ILA routers
- Basically an nvo3 control plane
- Initial development using BGP
- For scaling to to 100B objects may need more thought

BGP as control plane

- Why BGP
  - Reuse exsiting protocol seems attractive
  - BGP known to scale to a few million prefixes
  - Easy to extend, simple changes
- BGP ILA AFI
  - Locator value: 8 octets
  - Identifier(s): 8 octets

Comparison to ILNP

- ILA is IPv6 only
- ILA is transparent to transport layer
  - Symmetric address translation
  - Checksum neutral mapping
- UDP instead of ICMP for redirects

More comparison to ILNP

- Control plane not integrated
  - Leverage nvo3 control plane
  - We are working on BGP now
- Untranslated (ie. SIR) addresses routable
  - See topology
  - No requirements on DNS, ND
  - End host discovery by redirect

Alternatives considered in IPv6

- Use flow label for VNI
  - Non participating hosts won’t know this
  - Only 20 bits of information
  - Not covered by transport checksum
- Use extension headers, hold virtual address in EH for instance
  - Per 2460bis draft EHes can’t be added in flight
  - Not covered by transport checksum
  - Peformance, compatibility with network

Deployment steps

- IPv6 network needed
- Assign /64 to each host
  - Need to route to hosts based on /64
  - Configure DC routing hierarchy accordingly
- Deploy ILA routers
  - Initially assuming routers hold full table
  - ILA routers are assigned anycast SIR
  - They translate SIR to locators in forwarding
- Configure SIR prefix on hosts

ILA Identifier creation/registration

- Host (job scheduler, etc.) creates identifier
- Register {Identifier, Locator} in control plane, where locator is where object initially resides
- Control plane inform ILA routers of mapping
- Register name, SIR:ID in lookup service (DNS)
- Host connect to SIR:ID. ILA routes, redirects eliminate triangular routing

Status

- 4 I-Ds posted
- Data path integrated into Linux 4.1
- Canary testing (not migration though)
- Phase 1 deployment @FB
    - Assign /64 to every host
    - Task identifier generation
    - ILA router development

Questions?

Suggestions on how to proceed in IETF?

Thankyou!