# Serial Attached SCSI Architecture

by Rob Elliott

HP Industry Standard Servers

Server Storage Advanced Technology

elliott@hp.com    http://www.hp.com

30 June 2003

# Notice

- These slides are freely distributed by HP through the SCSI Trade Association (http://www.scsita.org)
- STA members are welcome to borrow any number of the slides (in whole or in part) for other presentations, provided credit is given to the SCSI Trade Association and HP
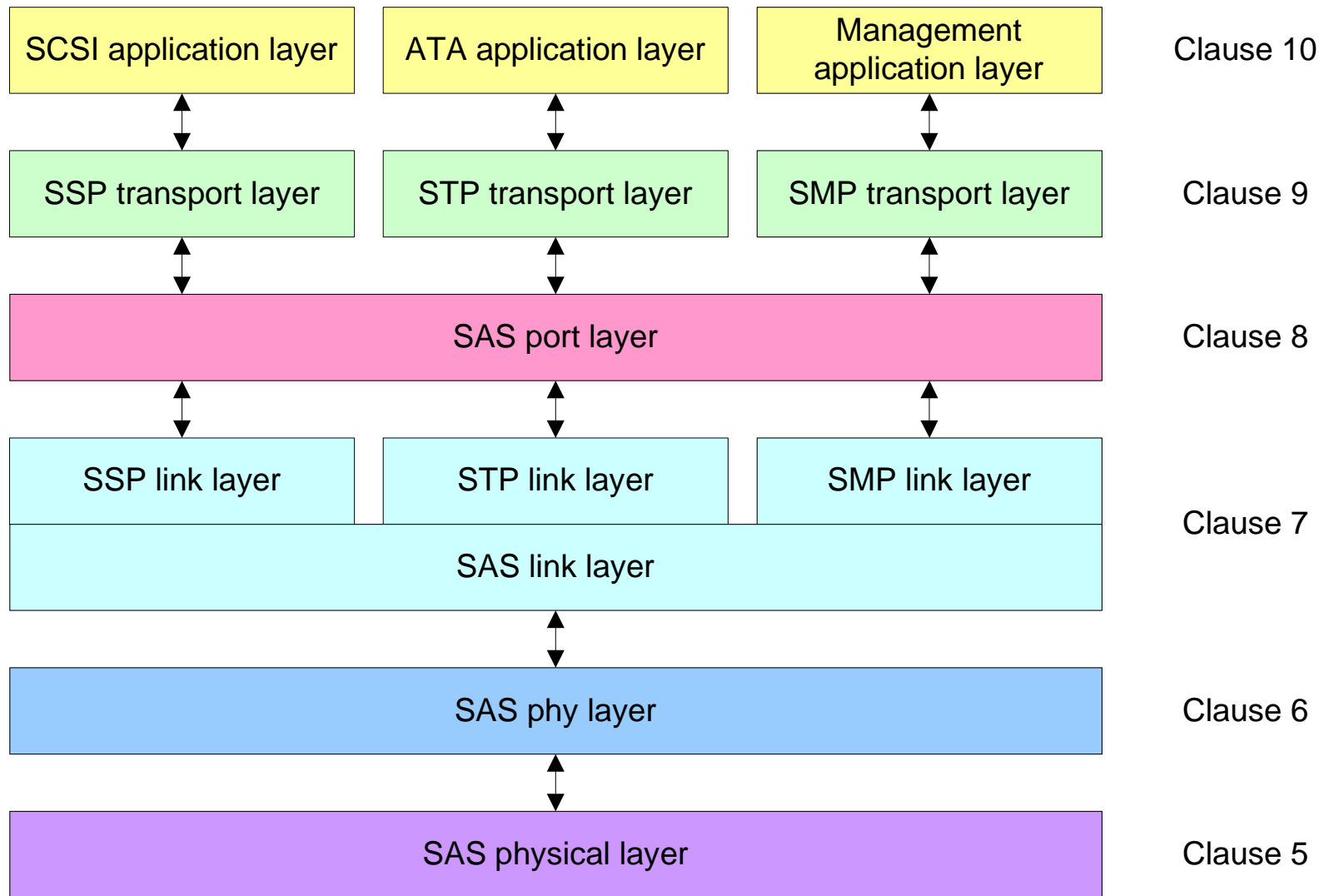- This compilation is © 2003 Hewlett-Packard Corporation

# SAS clause 4 – Architecture

- Standard layers

- SAS object model

- Physical links and phys

- Ports

- SAS devices

- Expander devices

- Domains

- Edge expander device set

- Pathways

- Connections

- SAS address

- Reset sequences

- State machines

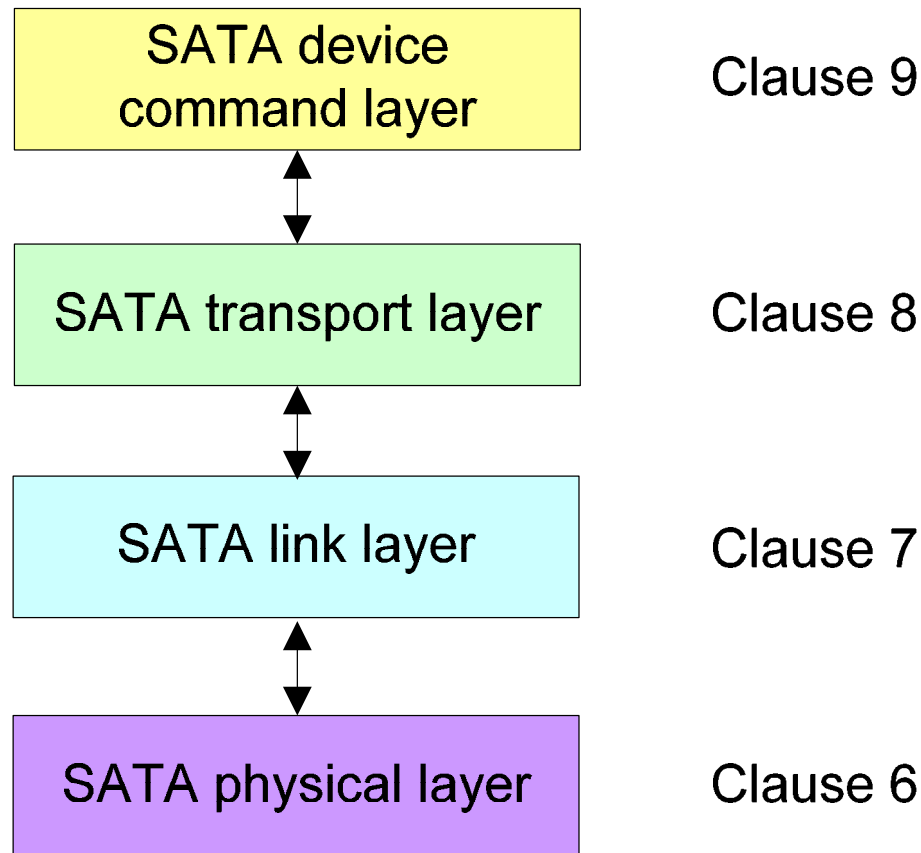- Transmit data path

- Expander device model

# SAS standard layering

| SCSI application layer | ATA application layer | Management application layer | Clause 10 |
|:---:|:---:|:---:|:---|
| SSP transport layer | STP transport layer | SMP transport layer | Clause 9 |
| SAS port layer | | | Clause 8 |
| SSP link layer | STP link layer | SMP link layer | Clause 7 |
| SAS link layer | | | |
| SAS phy layer | | | Clause 6 |
| SAS physical layer | | | Clause 5 |

# SATA 1.0a standard layering

- For SATA 1.0a from the private Serial ATA working group

| | |
|---|---|
| SATA device command layer | Clause 9 |
| SATA transport layer | Clause 8 |
| SATA link layer | Clause 7 |
| SATA physical layer | Clause 6 |

# ATA/ATAPI-7 standard layering

- For the public standard ATA/ATAPI-7
- Subject to change by T13 standards committee

| ATA command set | | Volume 1 Clause 4-6 |
|---|---|---|

| Parallel ATA transport layer | Volume 2 Clause 11 | | Serial ATA command layer | Volume 3 Clause 17 |
|---|---|---|---|---|
| Parallel ATA link layer | Volume 2 Clause 10, 12 | | Serial ATA transport layer | Volume 3 Clause 16 |
| Parallel ATA physical layer | Volume 2 Clause 7-9 | | Serial ATA link layer | Volume 3 Clause 15 |
| | | | Serial ATA physical layer | Volume 3 Clause 14 |
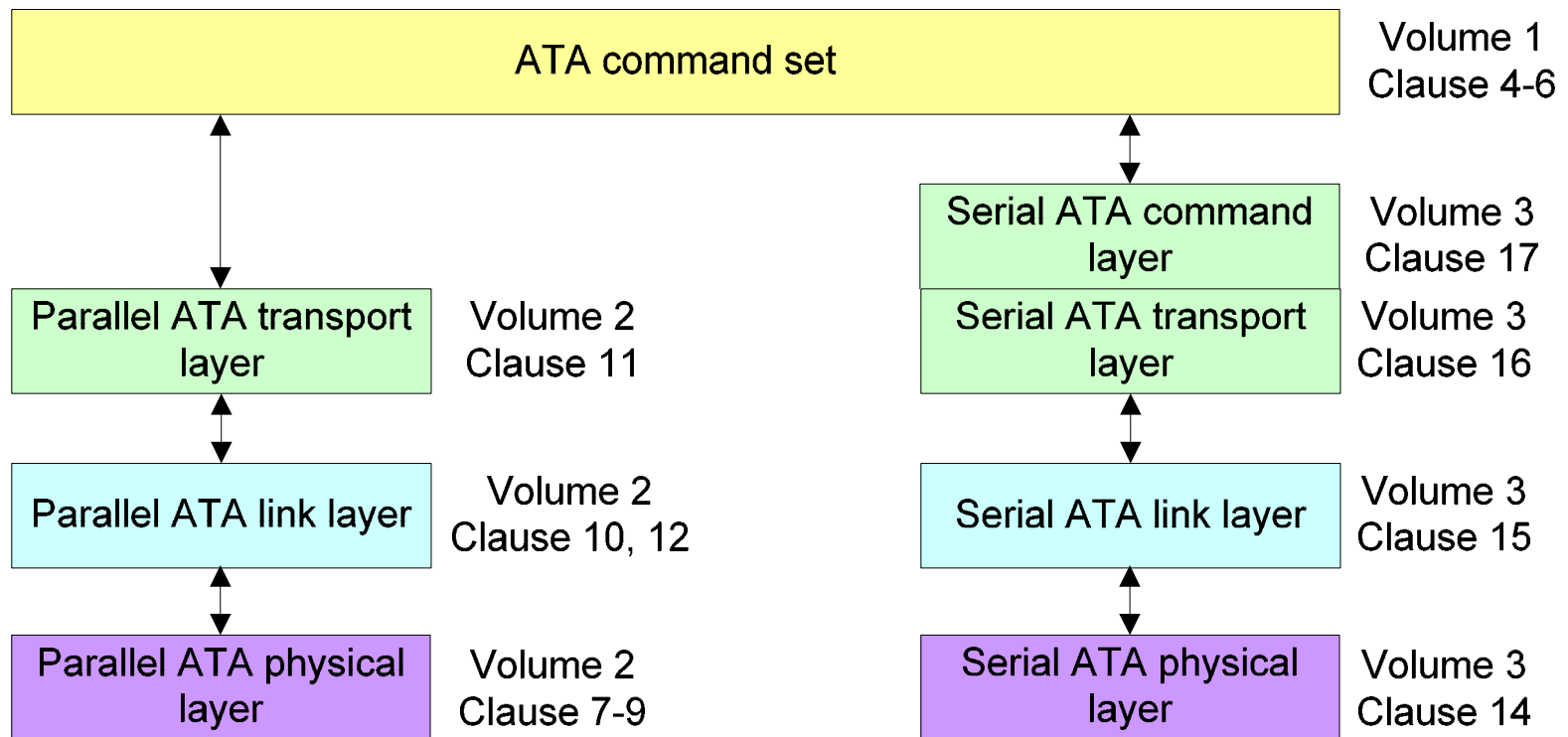
# SCSI and ATA terminology differences

| Common | SCSI | ATA |
|---|---|---|
| | SCSI device | <none> |
| HBA | SCSI initiator device | ATA host |
| Disk drive | SCSI target device | ATA device |
| | SCSI port | ATA port |
| | SCSI initiator port | ATA host port |
| | SCSI target port | ATA device port |

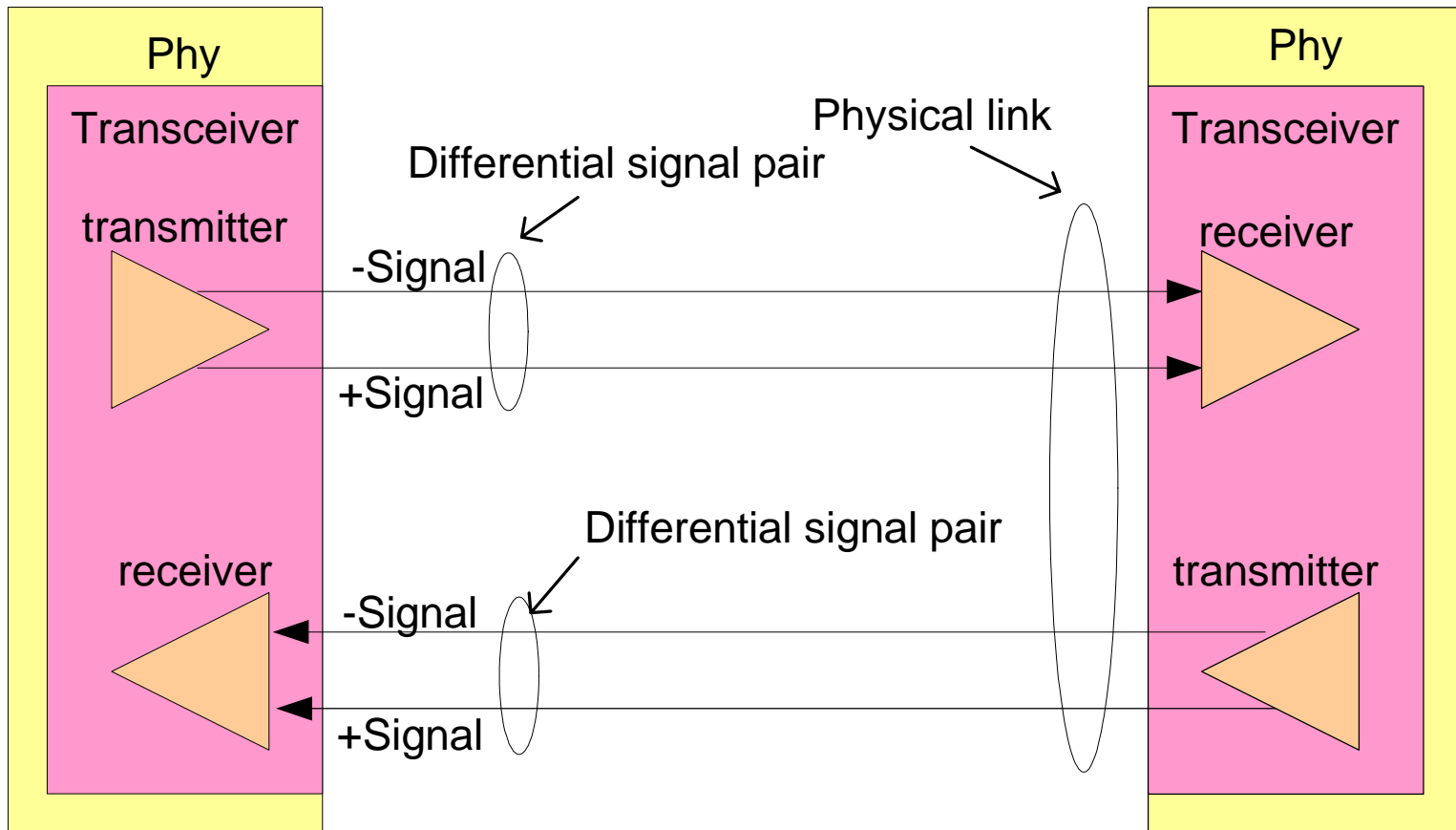| Common | SAS |
|---|---|
| | port |
| port | phy |

# SAS object model

- This figure describes classes of objects
- Diamonds mean "contains"
- Arrows mean "subclass of"
- Examples
  - SAS domain contains 1 or more SAS devices
  - SAS device contains one or more SAS ports
  - SAS port contains 1 to 128 phys
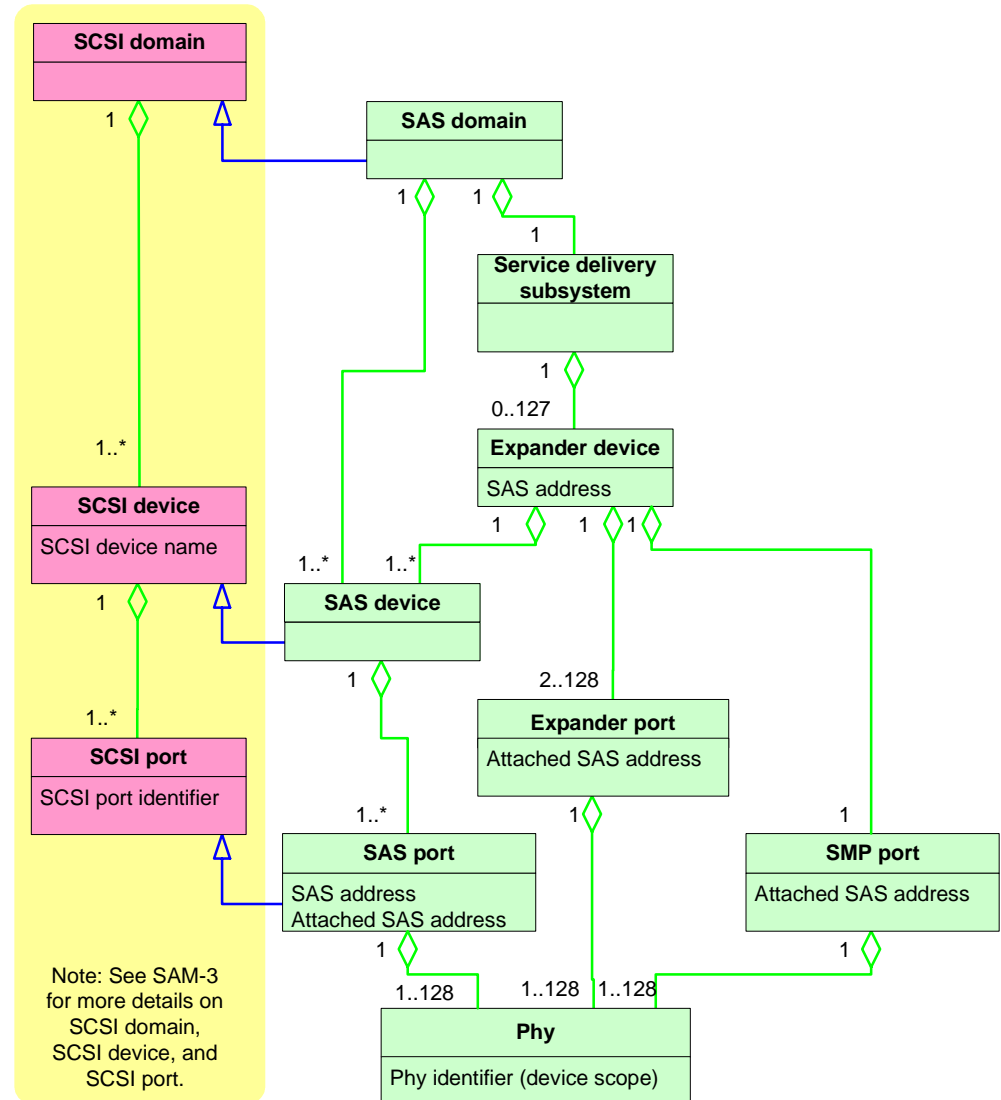
# Physical links and phys

- A phy contains one transceiver
- A physical link attaches two phys together

# Physical link rate

- Each direction runs 1.5 Gbps or 3.0 Gbps (150 MB/sec or 300 MB/sec)
  - Both directions use the same physical link rate
- Dual simplex (full duplex) operation – 600 MB/sec total bandwidth
- Example: peak bandwidth needs of an HBA with 8 phys
  - 2400 MB/sec half duplex, 4800 MB/sec full duplex
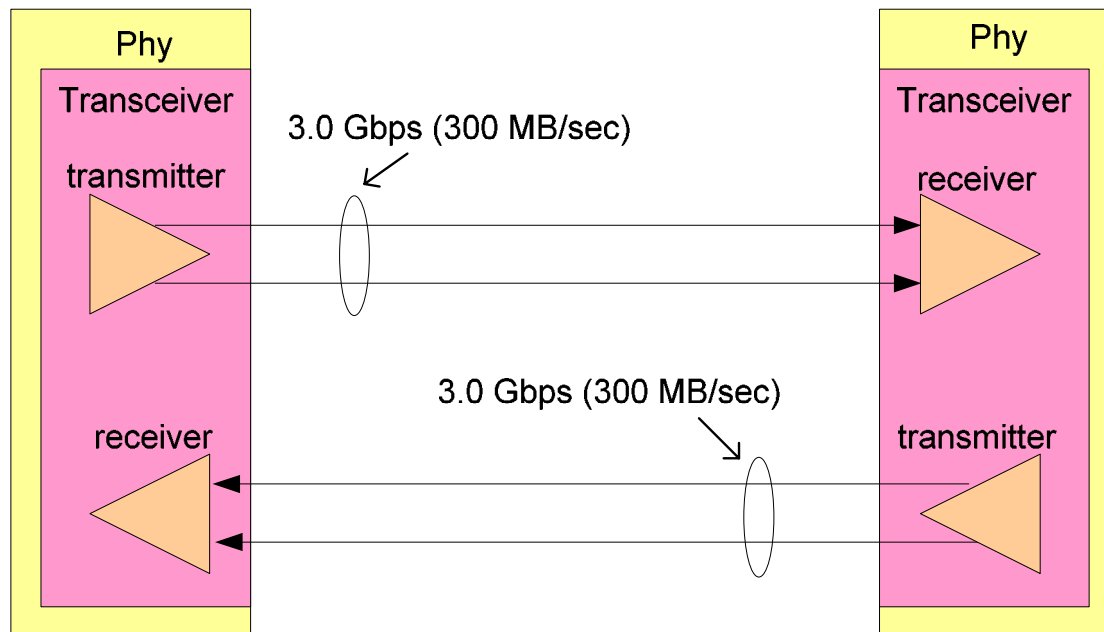
# Ports

- Ports contain phys
- An "expander port" is **not** a "SAS port"
- Each SAS port has a SAS address
- Ports are abstract
  - A set of phys with matching SAS addresses attached to another set of phys with matching SAS addresses
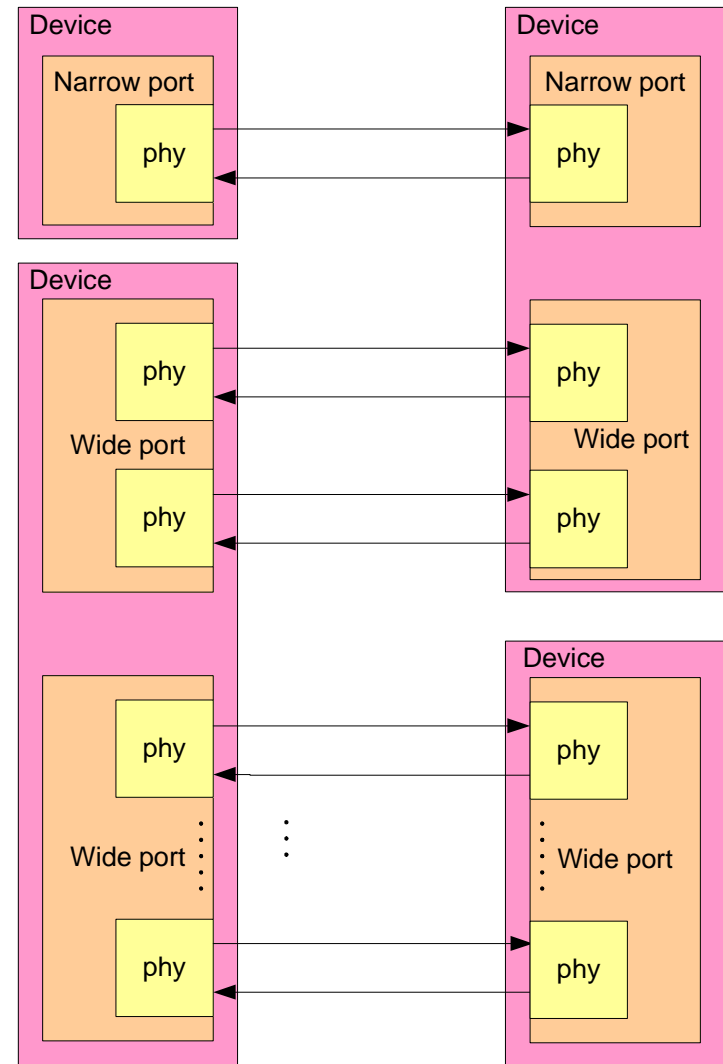  - Determined at initialization time
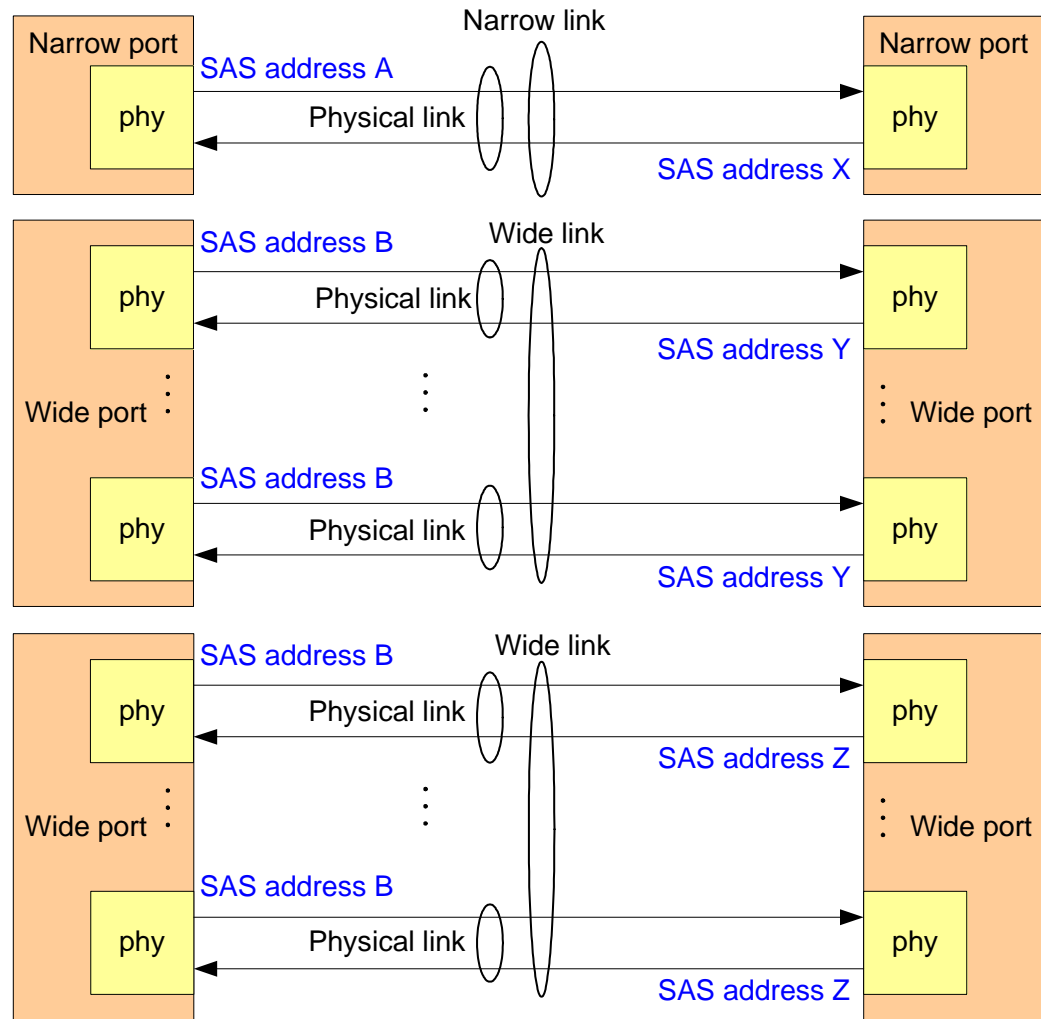


Each horizontal line represents a differential signal pair

# SAS devices

- SAS devices contain ports
- An "expander device" is **not** a "SAS device"
- Each phy in a SAS device has a phy identifier unique within that device

Each horizontal line represents a differential signal pair

# End devices

- End device is a SAS device that is not an expander device
- Sample end devices
  - HBA – 8 phys
    - One SAS address for all 8 phys
      - Potentially all one (very) wide port
    - One SAS address for 4 phys, another SAS address for 4 phys
      - Guarantees at least two ports
      - Good match for 4-wide connectors
    - Eight SAS addresses
  - Disk drive - 2 phys
    - Separate SAS address for each phy
      - Guarantees two ports
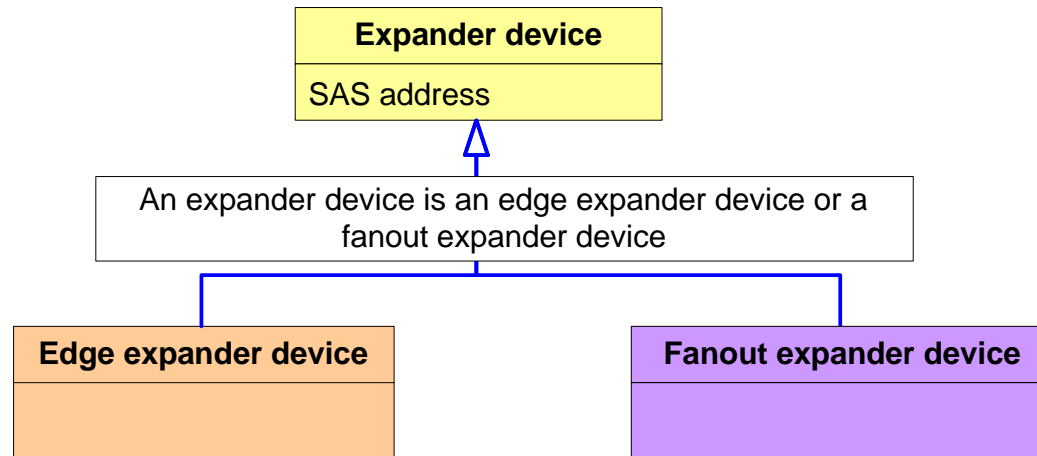      - Never a wide port

# Expander devices

- Expander device contains expander ports
- May contain SAS devices too (e.g. for enclosure management)
- Each expander device has a SAS address
- Each expander phy has a phy identifier unique within that expander device
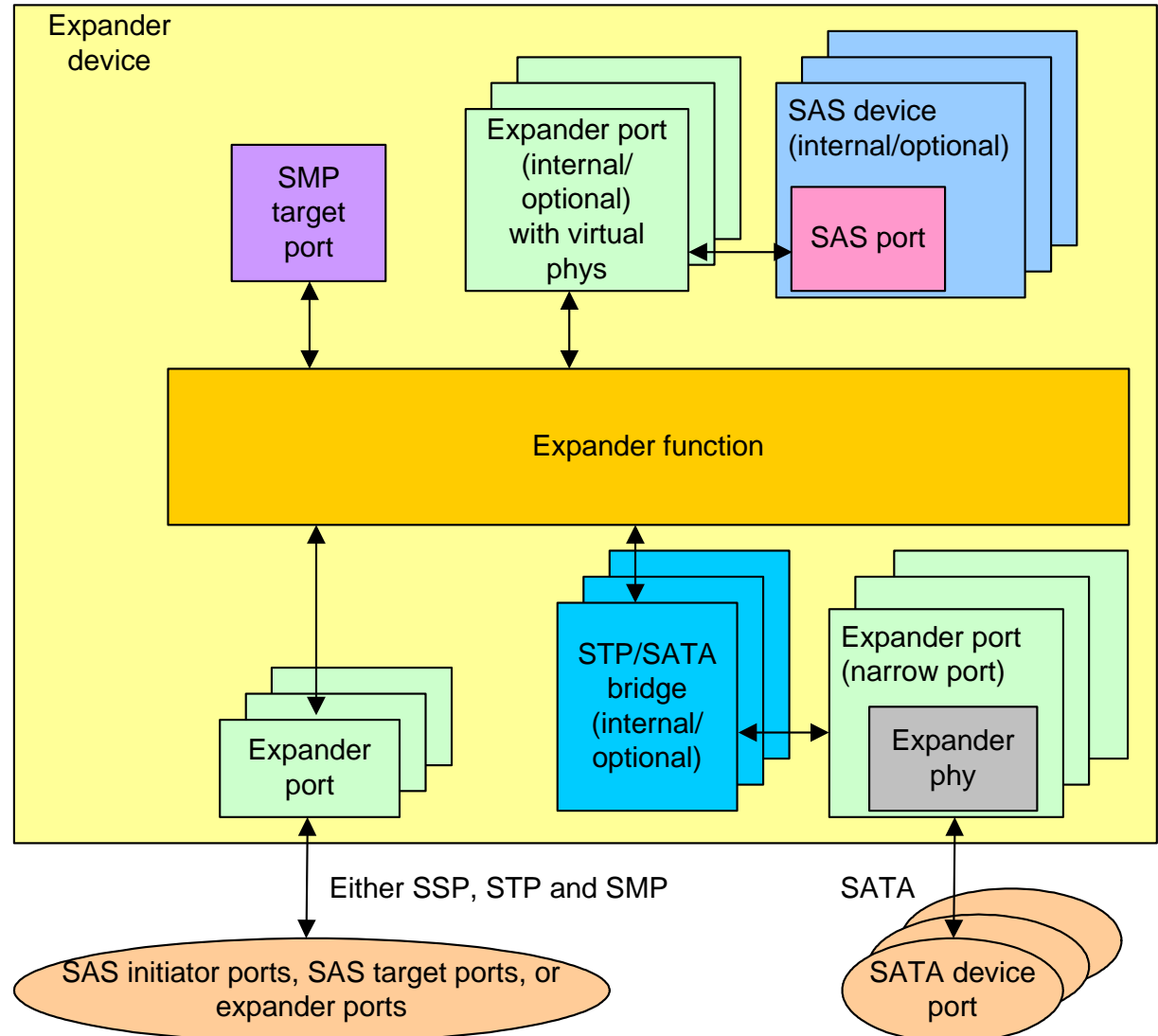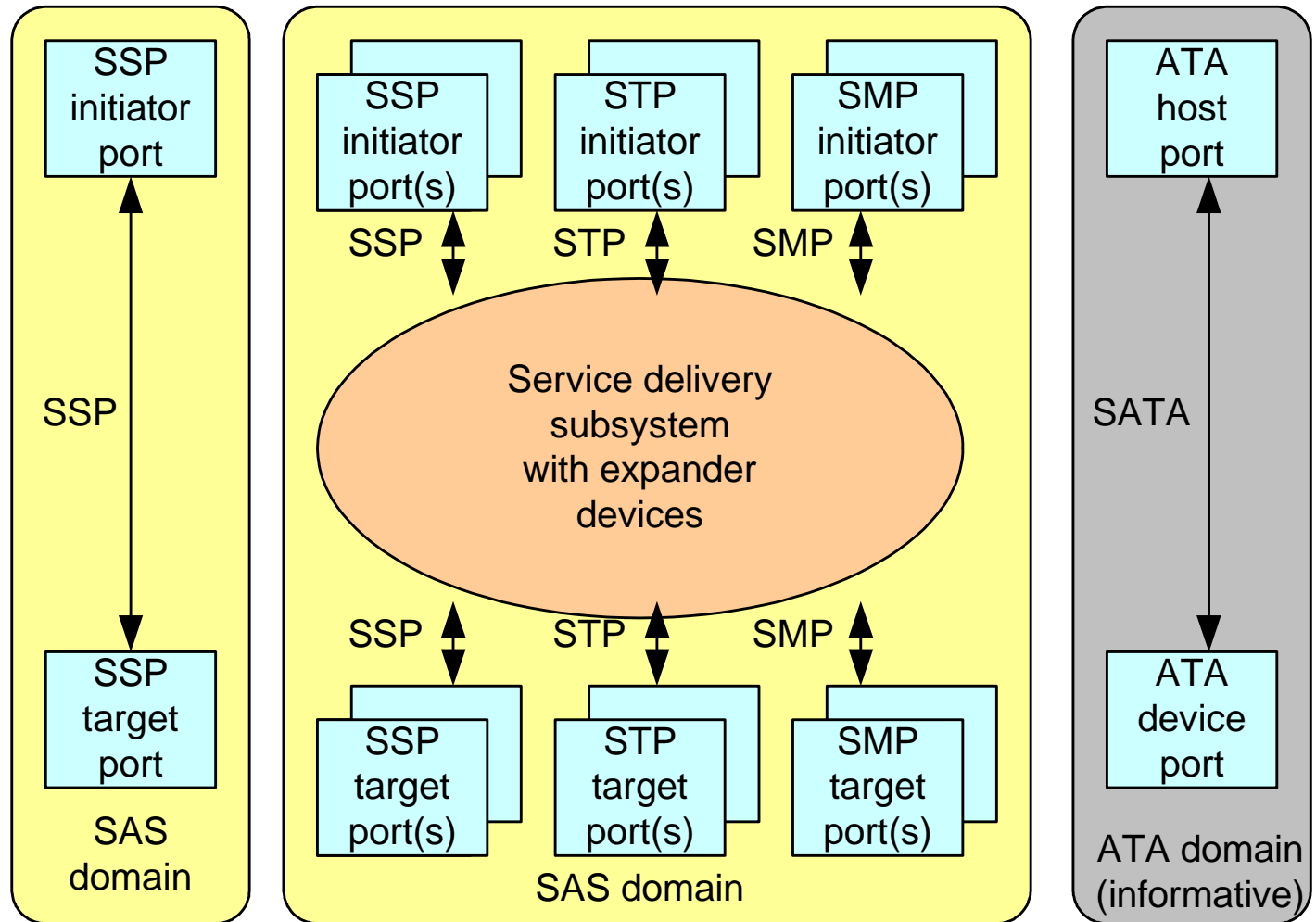
# Expander device types – edge vs. fanout

```
        ┌─────────────────────────┐
        │    Expander device      │
        ├─────────────────────────┤
        │ SAS address             │
        └─────────────────────────┘
                    △
        ┌─────────────────────────────────────────────┐
        │ An expander device is an edge expander device or a │
        │           fanout expander device              │
        └─────────────────────────────────────────────┘
          │                                      │
┌──────────────────────────┐      ┌──────────────────────────┐
│  Edge expander device    │      │  Fanout expander device  │
├──────────────────────────┤      ├──────────────────────────┤
│                          │      │                          │
└──────────────────────────┘      └──────────────────────────┘
```

- Edge expander device
  - Always part of an "edge expander device set"
  - May perform subtractive routing
- Fanout expander device
  - Never does subtractive routing
  - Usually supports larger tables for table routing
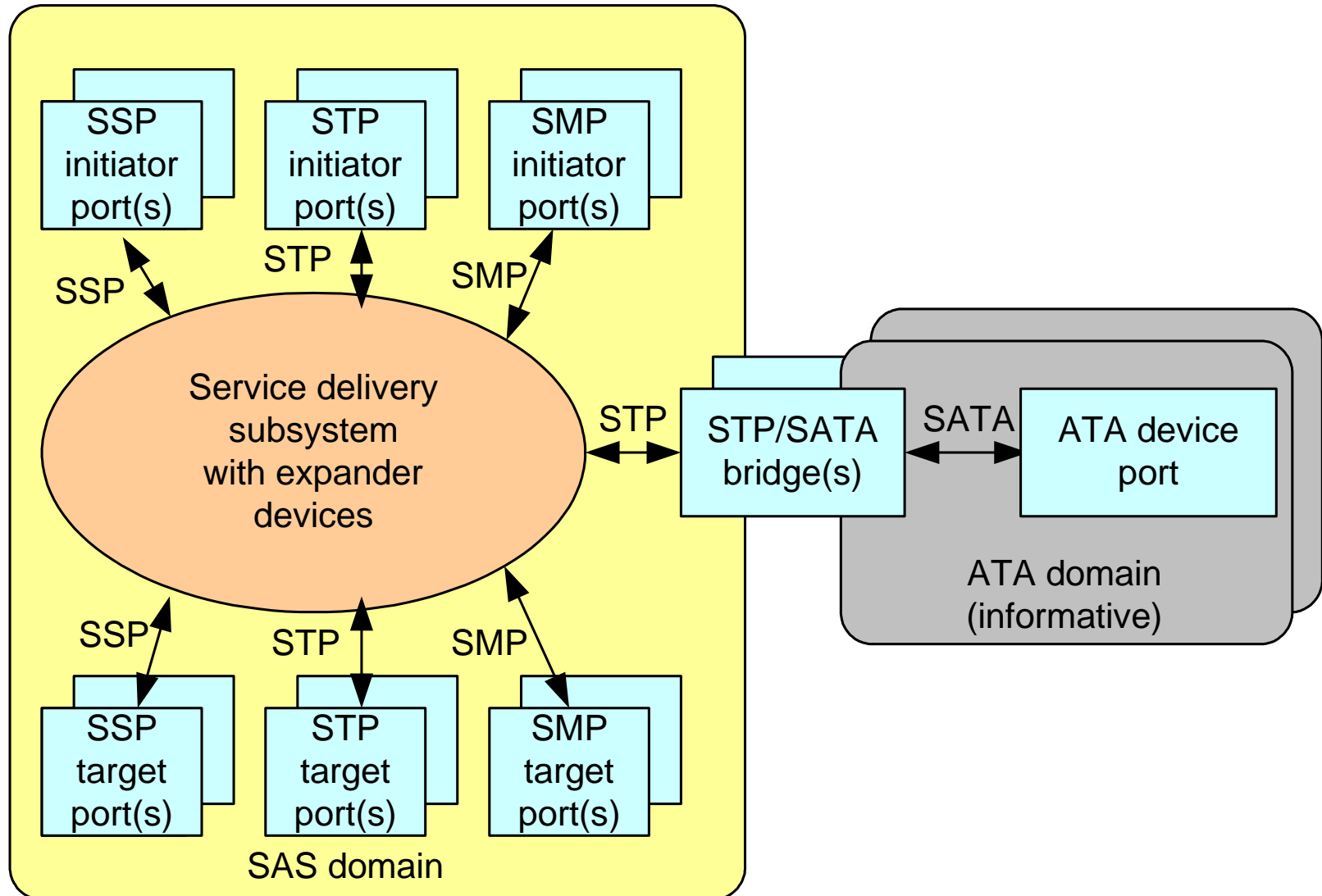- Topologies described later

# Domains

- A simple SAS domain contains SAS devices and expander devices
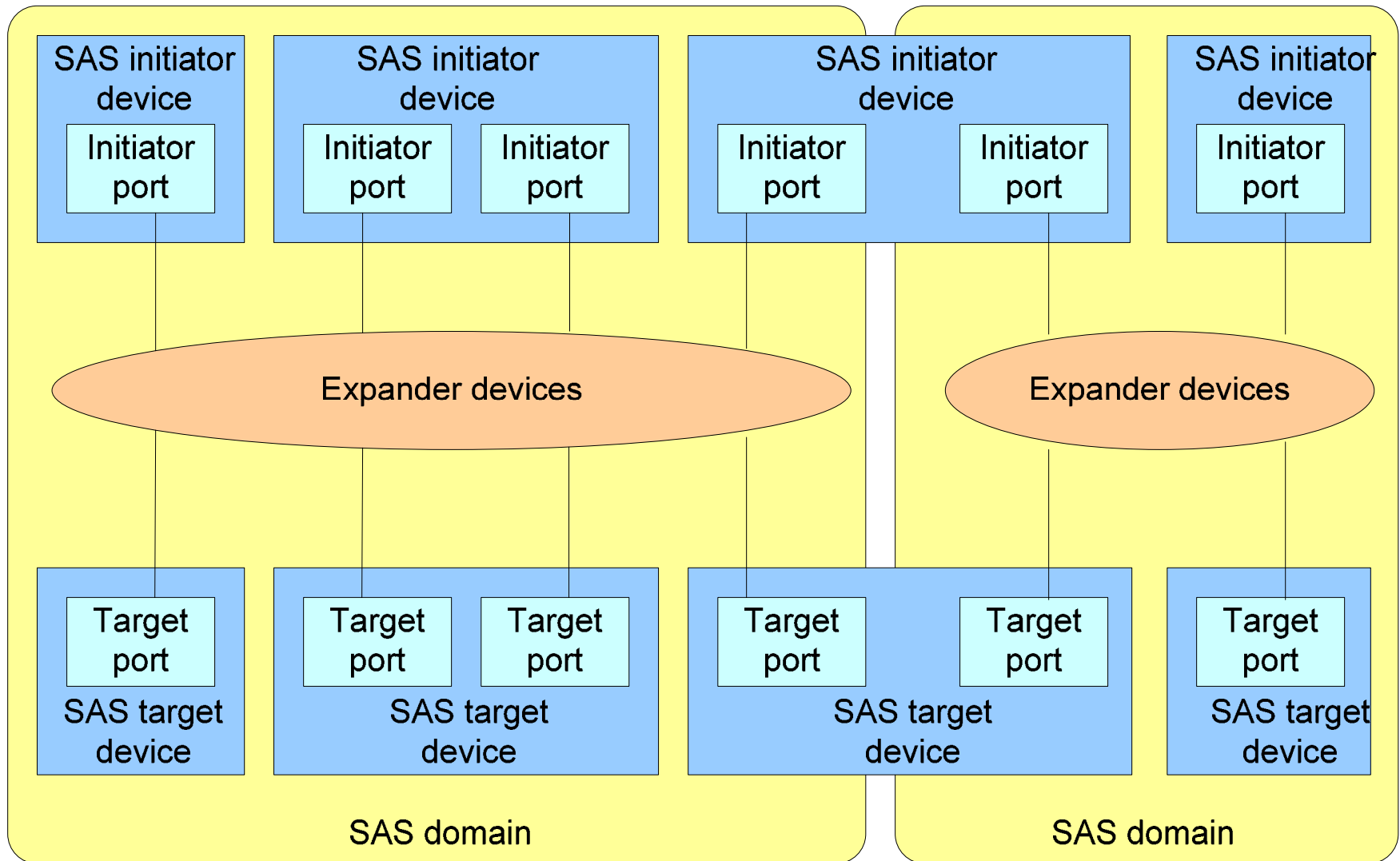- An ATA domain contains a SATA host and a SATA device



Note: When expander devices are present, SAS target ports may be located in SAS devices contained in expander devices.

# SAS domain bridged to ATA domains

# SAS devices in multiple SAS domains



SAS initiator device
- Initiator port

SAS initiator device
- Initiator port
- Initiator port

SAS initiator device
- Initiator port
- Initiator port

SAS initiator device
- Initiator port

Expander devices

Expander devices

Target port — SAS target device

Target port — Target port — SAS target device

Target port — Target port — SAS target device

Target port — SAS target device
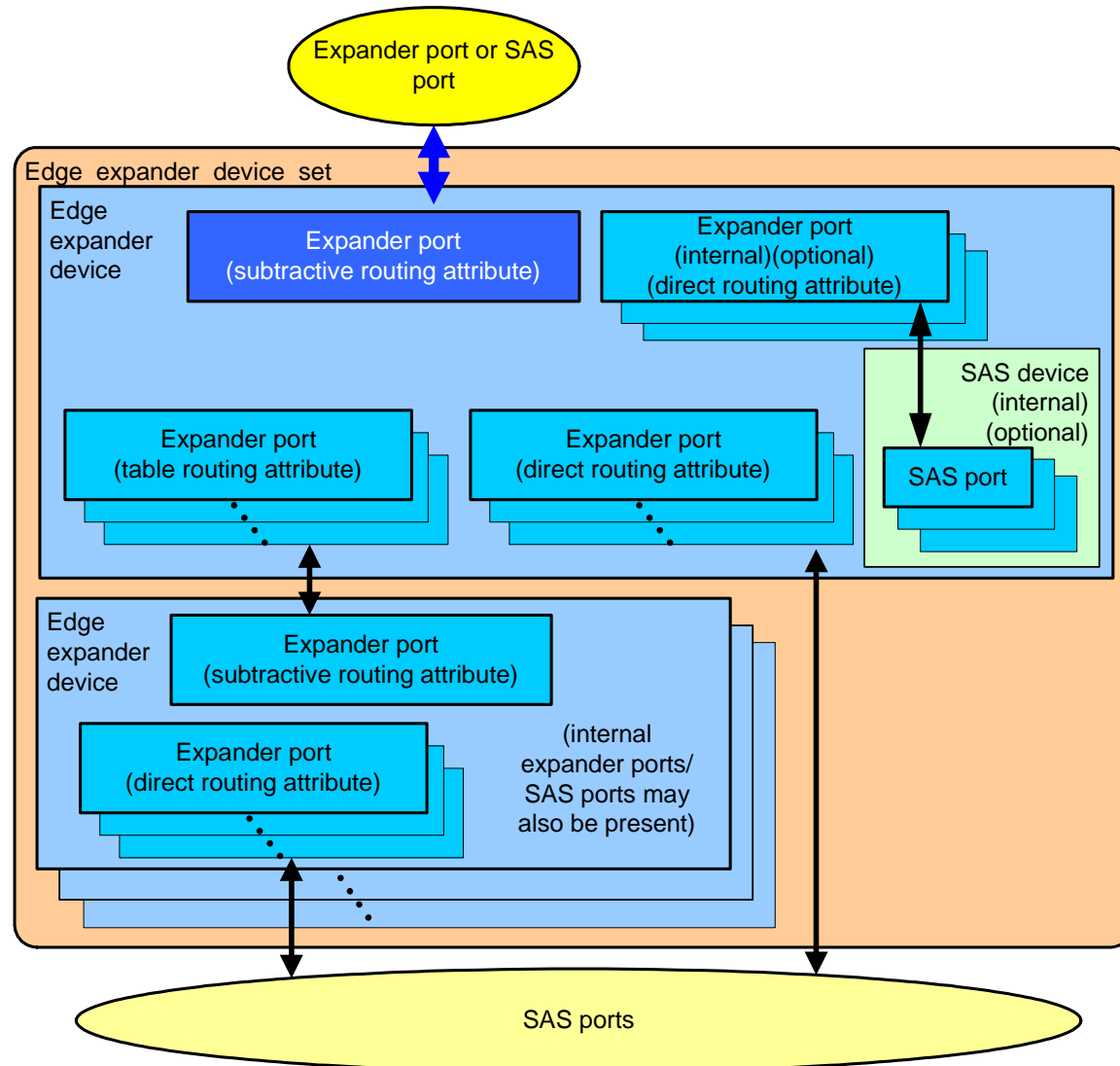
SAS domain

SAS domain

# Edge expander device set

- Set of edge expander devices
- 128 SAS addresses per set
- Typically bounded by a subtractive port (to a fanout expander device, or to another edge expander device set)
- Edge expander devices uses table routing and direct routing "downstream" and subtractive routing "upstream"
- Wide links between expanders are allowed
- No loops

SAS architecture
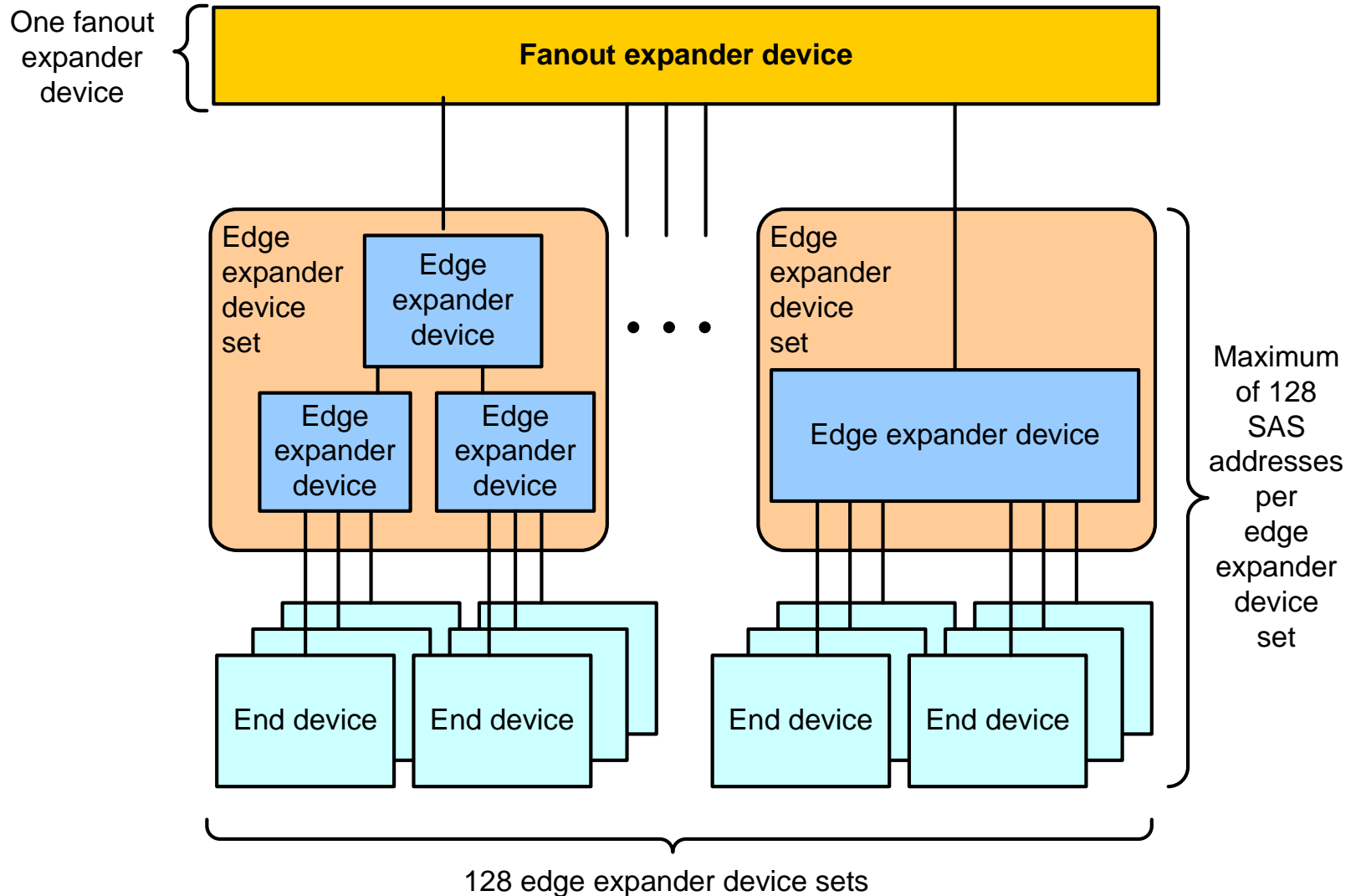
# Edge expander device set diagram

# Expander topologies

- Maximum of one fanout expander device in a SAS domain
- If no fanout expander, maximum of two edge expander device sets (attached via subtractive decode ports)
- End devices may be attached at any level
  - Directly to fanout expander device
  - Any level edge expander device
- Wide links possible between any two devices
- No loops
- No multiple paths

# Edge expander device set and fanout expander device



One fanout expander device

Fanout expander device

Edge expander device set

Edge expander device

Edge expander device

Edge expander device

• • •

Edge expander device set

Edge expander device

End device

End device

End device

End device

Maximum of 128 SAS addresses per edge expander device set

128 edge expander device sets

# Two edge expander device sets



The root edge expander device in each edge expander device set uses the subtractive routing method on the expander phys attached to its peer.

Upstream phys use the subtractive routing method; downstream phys use table routing method or direct routing methods.

Edge expander device set

Edge expander device

Edge expander device

Edge expander device

Edge expander device set

Edge expander device

Edge expander device

Edge expander device

End device

End device

End device

End device

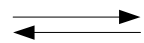End device

Two edge expander device sets
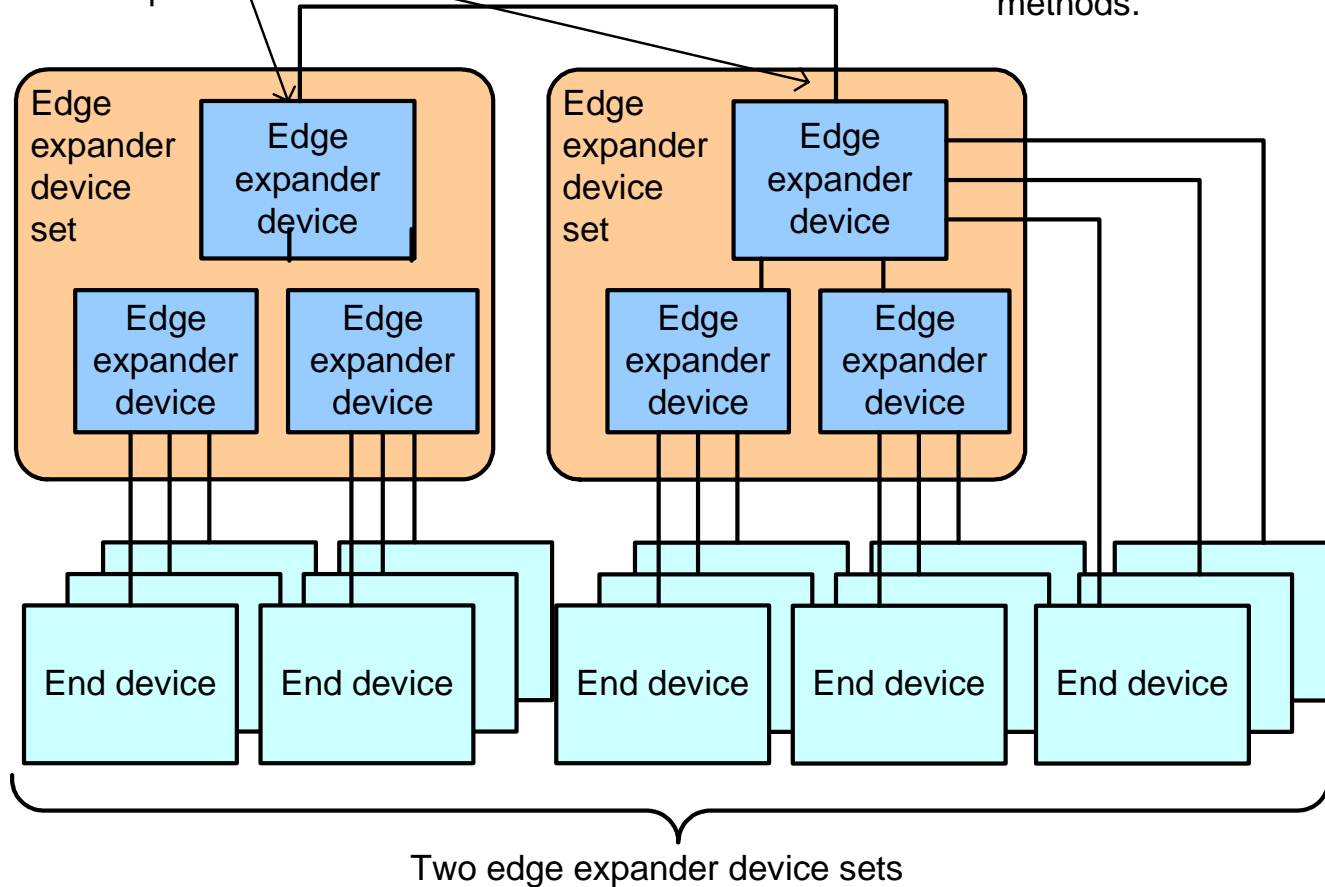
# Pathways

- Potential pathway = set of physical links between an initiator phy and a target phy
- Pathway = set of physical links used by a connection



Key:
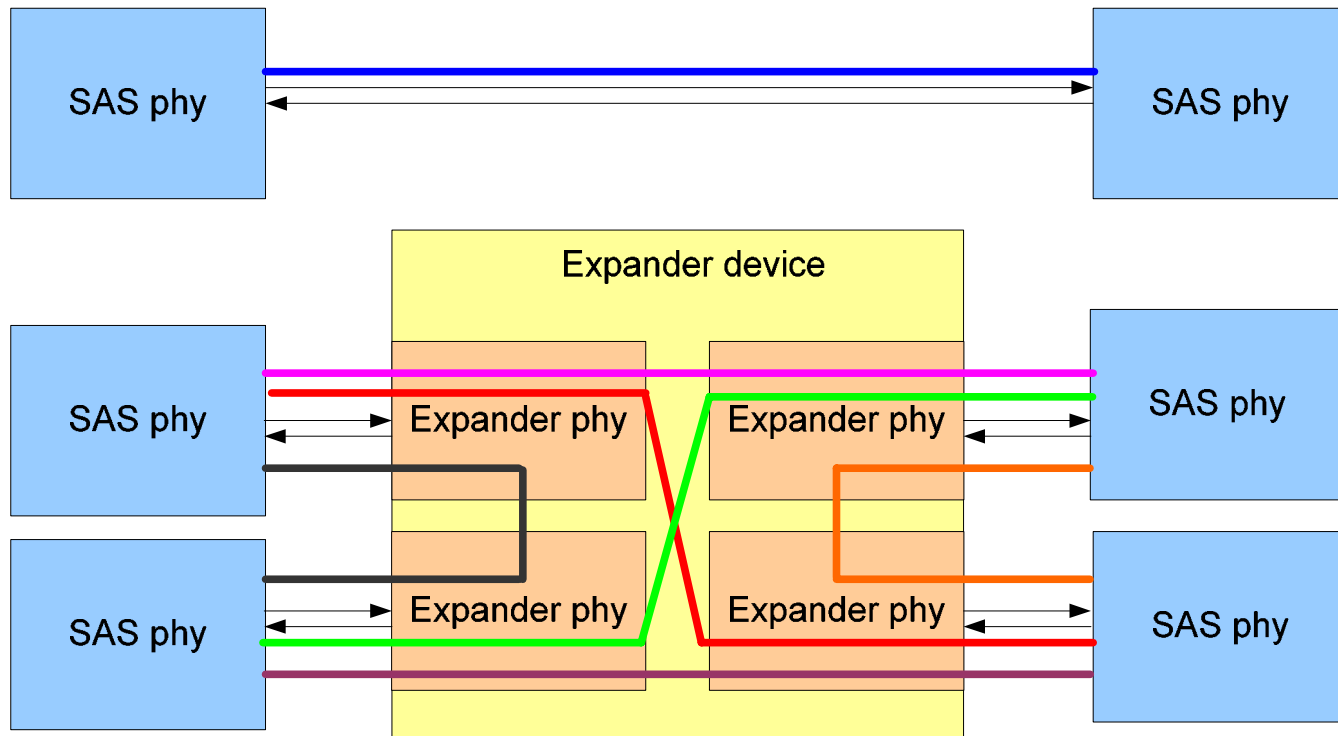
Single physical link          Potential Pathway

# Connections

- Connection = temporary association between an initiator phy and a target phy
    - Source phy transmits an OPEN address frame
        - Contains a destination SAS address
    - Expanders route it to a matching destination phy
    - Destination phy replies with an OPEN_ACCEPT primitive
    - Connection is established
    - Both sides exchange CLOSE primitives to close
- Connections are addressed to ports but established phy-to-phy
- N-wide ports may establish N connections at a time (to up to N other ports)
- Wide ports may establish multiple connections to other wide ports simultaneously

# Connection examples
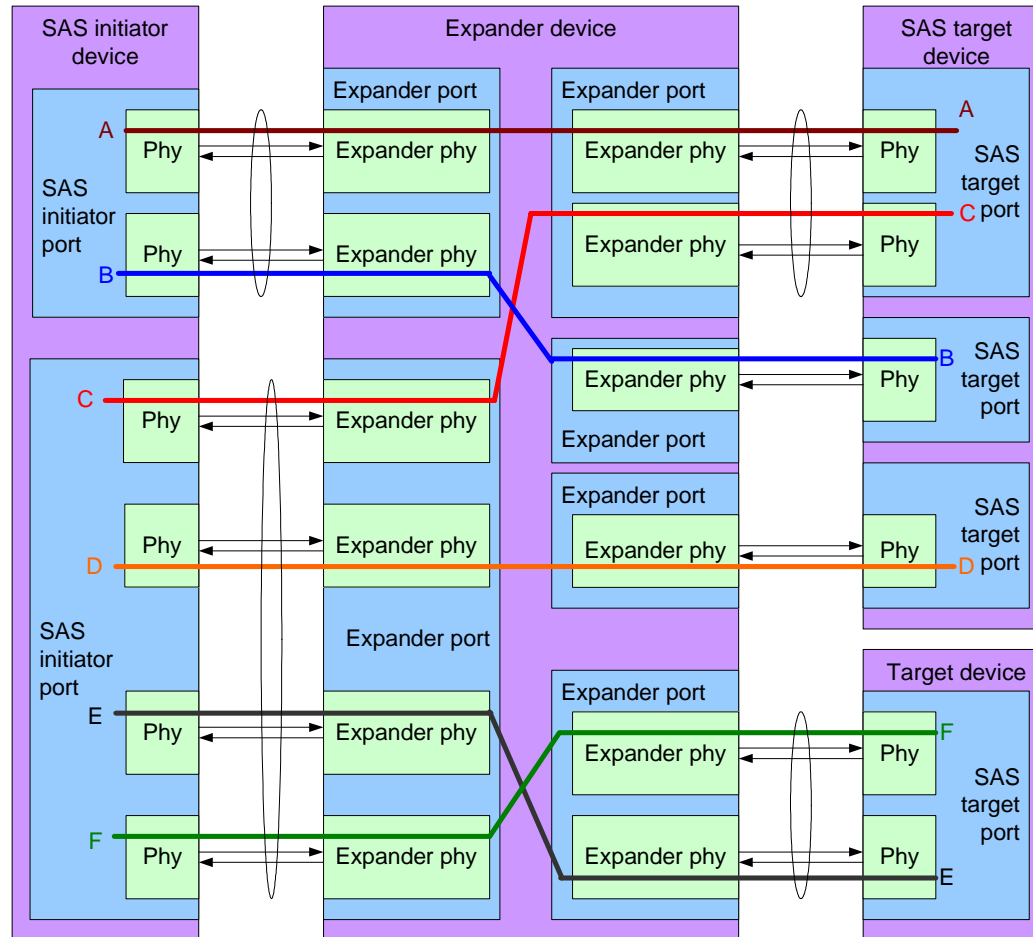


**Key:** ───▶ Single physical link  ◯ Wide link  X───X Connection

Notes: The expander device has a unique SAS address. Each SAS initiator port and SAS target port has a unique SAS address. Connections E and F represent a wide SAS initiator port with two simultaneous connections to a wide SAS target port.

# Connection rules

- Connections are addressed to SAS ports but are established from phy to phy
- Wide ports may establish multiple connections at a time (to up to one per phy) to different destinations
- Wide ports may establish multiple connections to other wide ports simultaneously (wide initiator port to wide target port)
  - SAS disk drives will offer two narrow ports
  - Only HBAs and RAID controllers will offer wide ports

# Connection rate

- Connection runs at 1.5 Gbps or 3.0 Gbps
- Connection rate <= physical link rate
- If the connection rate is slower than the physical link rate, rate matching is used
  - e.g. 1.5 Gbps connection rate over a 3.0 Gbps physical link
  - Inserts ALIGN primitive every other dword to slow down the effective throughput
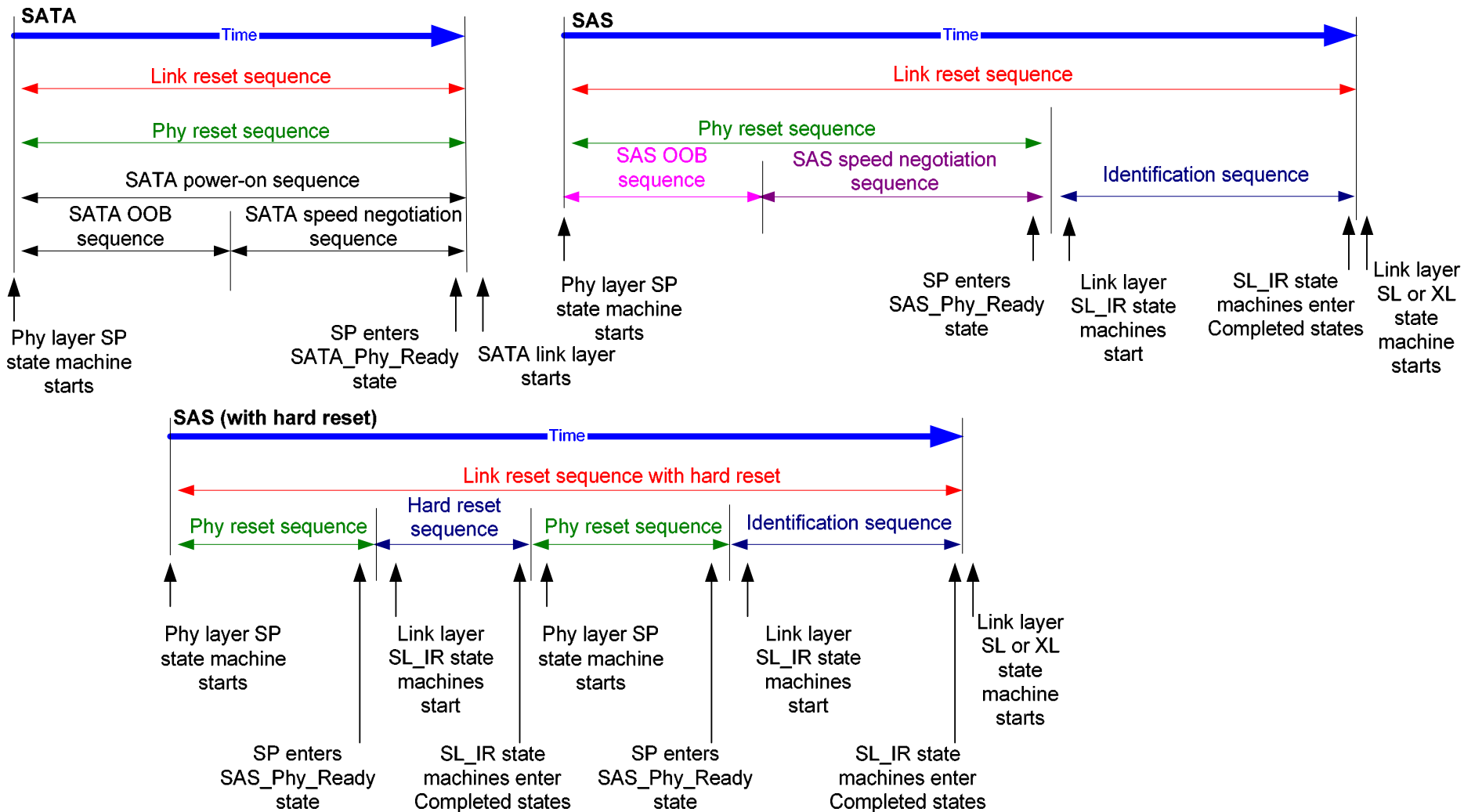  - Lets 3.0 Gbps initiators talk to 1.5 Gbps targets through expanders

# SAS address

- Each SAS port and expander device has a worldwide unique 64-bit SAS address
- Same namespace as the Fibre Channel Port_Name

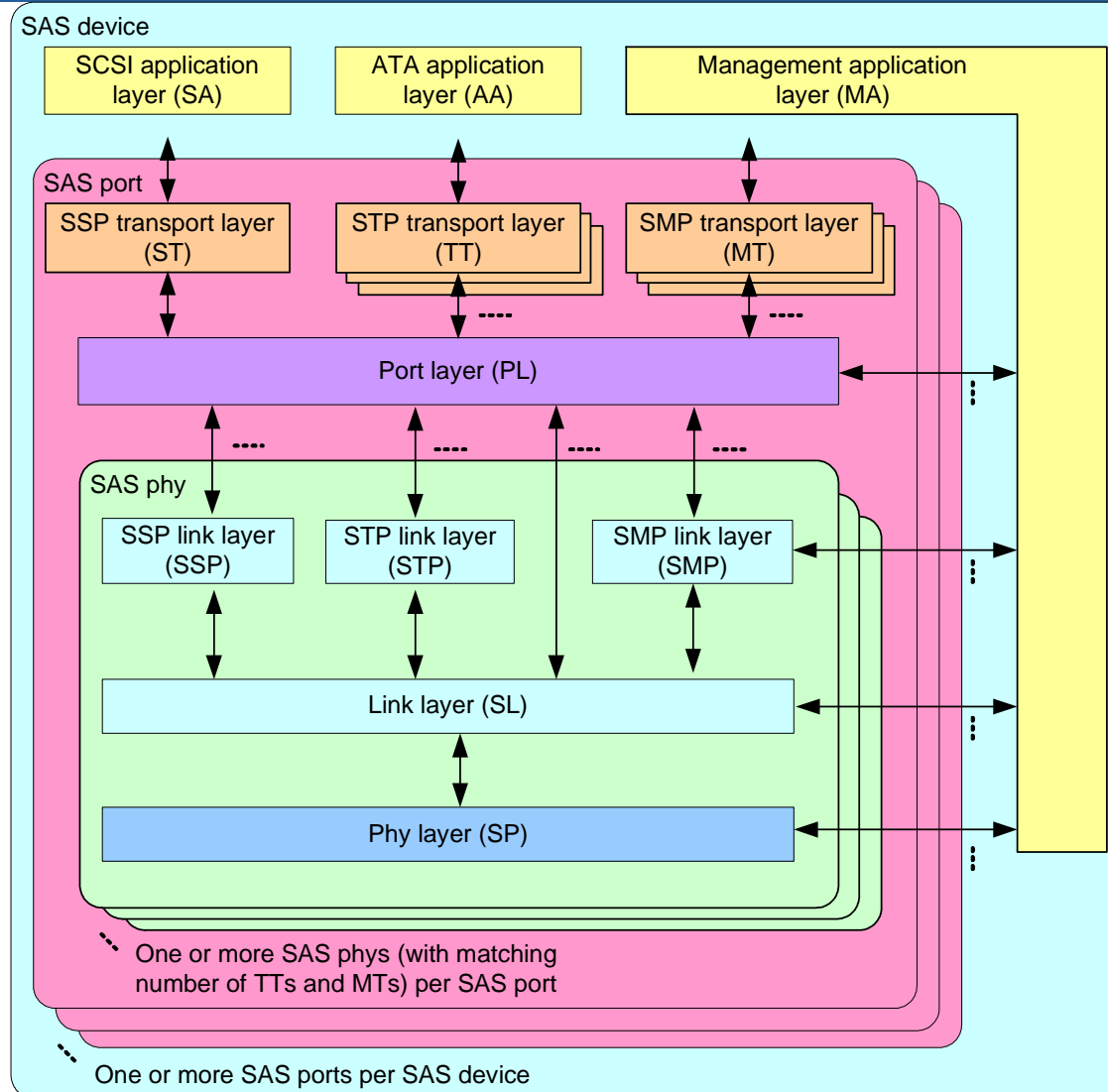| Byte\Bit | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|----------|---|---|---|---|---|---|---|---|
| 0 | NAA (5h) | | | | | | | |
| 1 | IEEE Company ID (24 bits) | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |
| 4 | Vendor-Specific Identifier (40 bits) | | | | | | | |
| 5 | | | | | | | | |
| 6 | | | | | | | | |
| 7 | | | | | | | | |

# Reset sequences

# State machines

- SAS uses state machines for detailed definition of rules
- Intended to reduce interoperability issues from reading vague text descriptions (e.g. that haunted parallel SCSI)
- SAS device state machines differ from expander device state machines
- SAS state machines do not provide:
  - Detailed integration of SATA state machines into SAS
  - Detailed implementation of expander functionality
  - Hardware-implementable descriptions (zero-time states, "state machines" with only one state, state transitions passing arguments, and other atrocities)

# State machines for SAS devices

- SAS device object contains application layer state machines (and SAS port objects)
- SAS port object contains transport layer and port layer state machines (and SAS phy objects)
- SAS phy object contains link layer and phy layer state machines



SAS device

| SCSI application layer (SA) | ATA application layer (AA) | Management application layer (MA) |

SAS port

SSP transport layer (ST) | STP transport layer (TT) | SMP transport layer (MT)

Port layer (PL)

SAS phy

SSP link layer (SSP) | STP link layer (STP) | SMP link layer (SMP)

Link layer (SL)

Phy layer (SP)

One or more SAS phys (with matching number of TTs and MTs) per SAS port
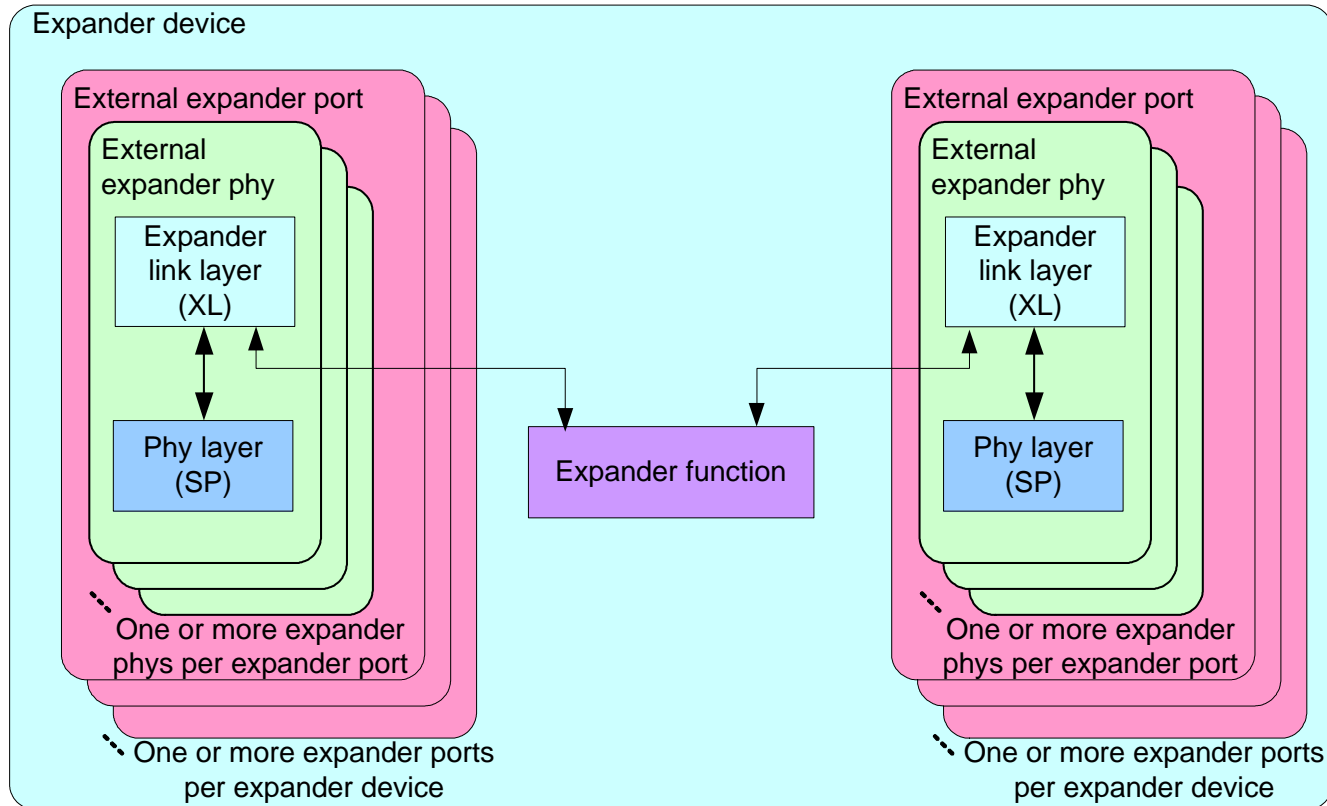
One or more SAS ports per SAS device

# State machines for expander devices

- Expander phy object contains link layer and phy layer state machines
- Expander port object contains expander phys
- Expander function is architecturally at a lower level than the link layer

# Transmit data path in SAS phy
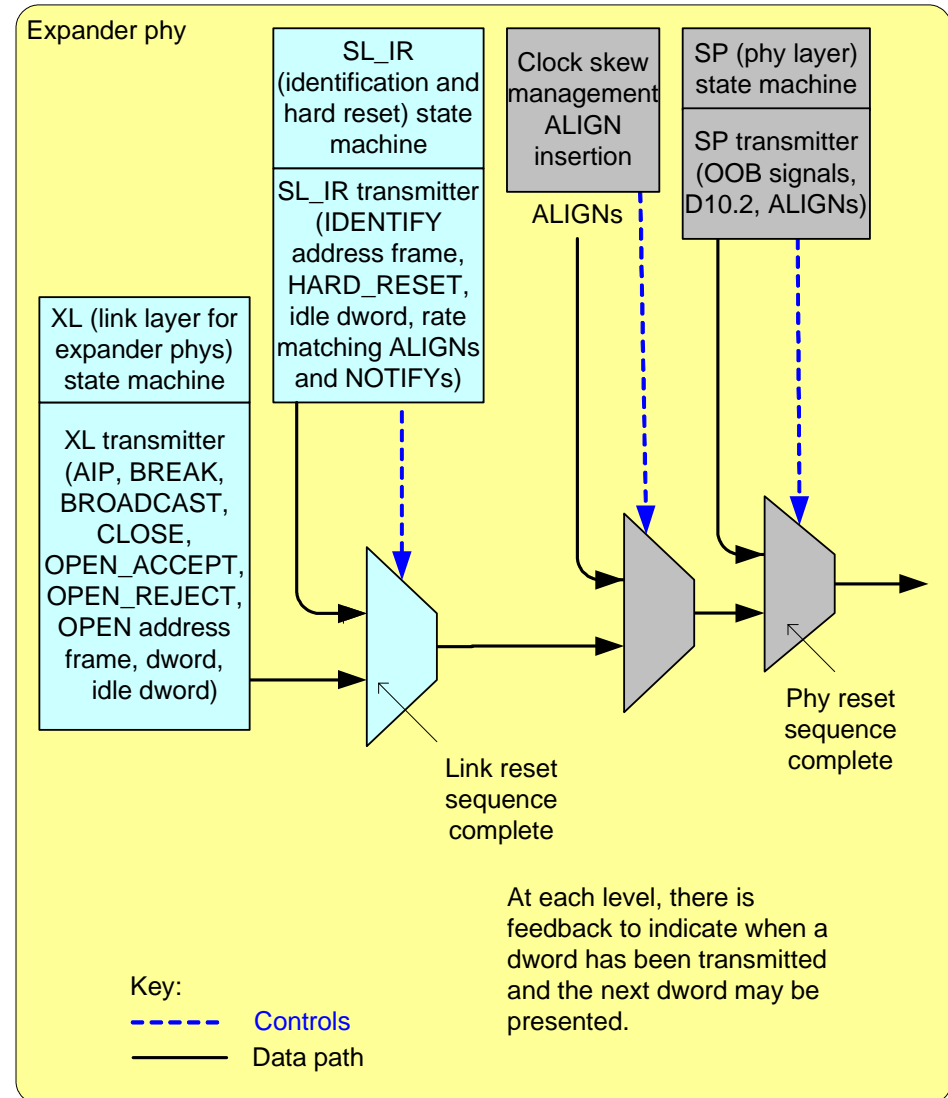
- SP transmits dwords during phy reset sequence
- SL_IR transmits dwords after phy reset sequence
- SL transmits dwords after identification sequence, outside connections
- SSP, STP, or SMP link layer transmits dwords during connections
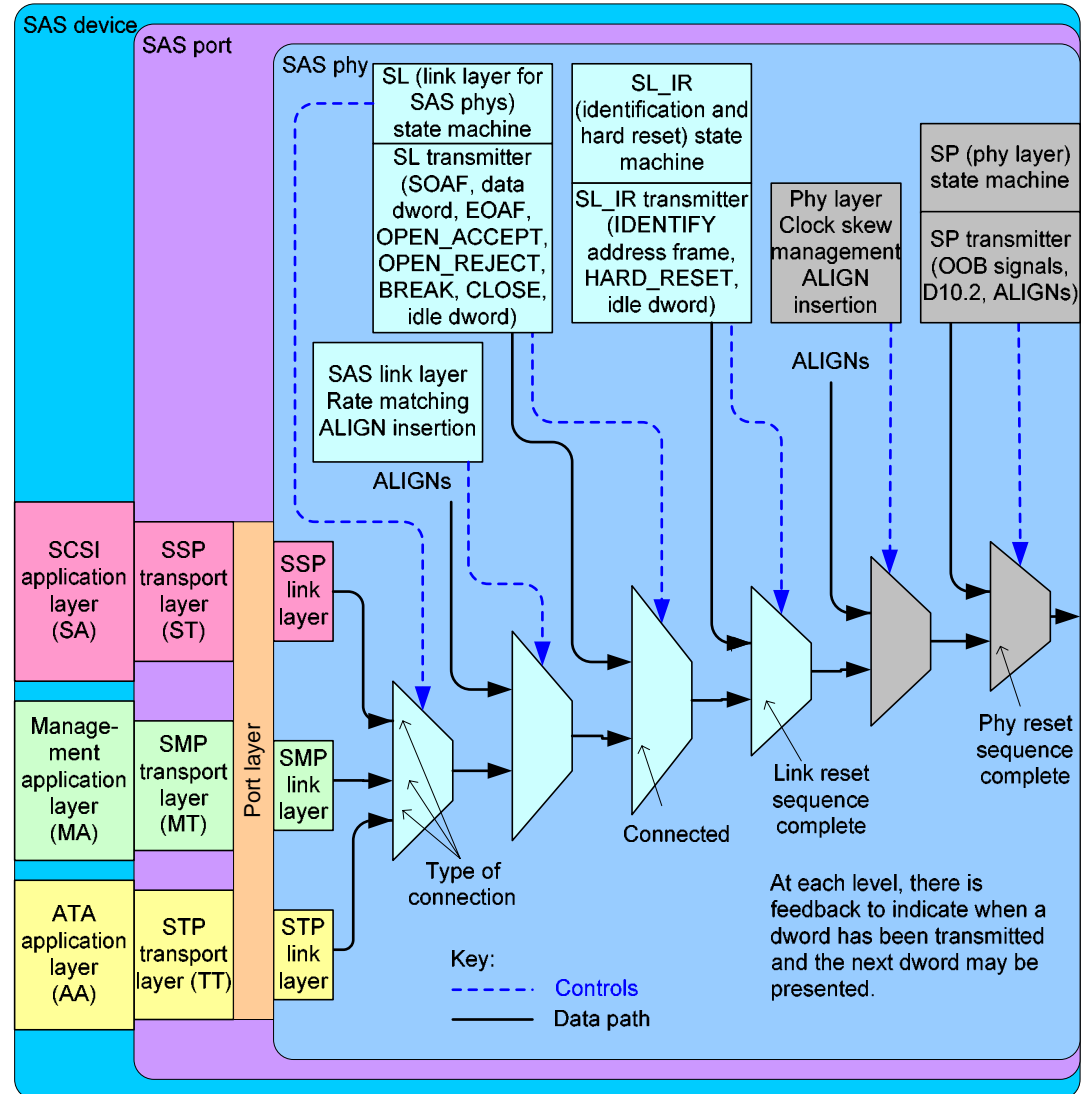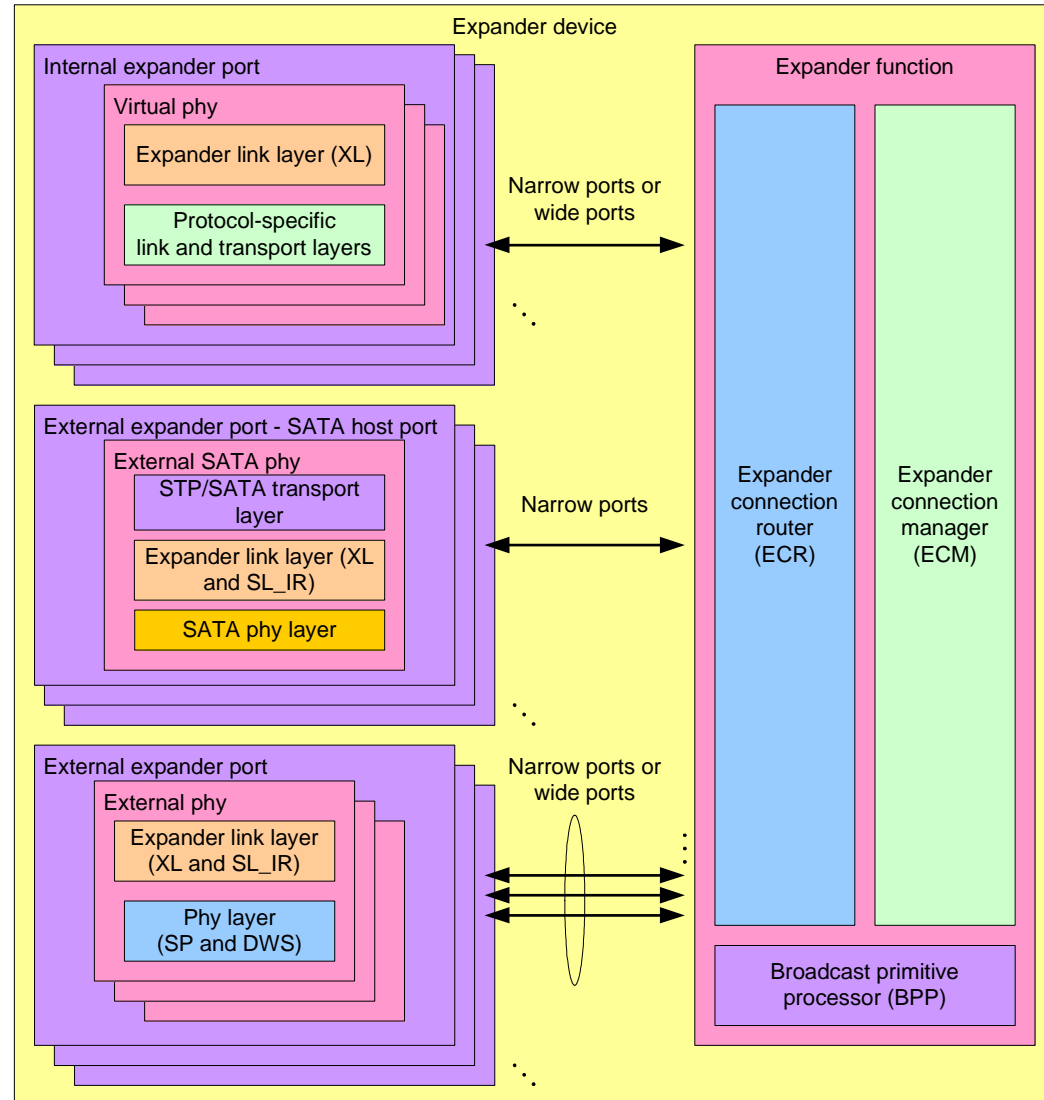
# Transmit data path in expander phy

- SP transmits dwords during phy reset sequence
- SL_IR transmits dwords after phy reset sequence
- XL transmits dwords after identification sequence



Expander phy

SL_IR (identification and hard reset) state machine

SL_IR transmitter (IDENTIFY address frame, HARD_RESET, idle dword, rate matching ALIGNs and NOTIFYs)

Clock skew management ALIGN insertion

ALIGNs

SP (phy layer) state machine

SP transmitter (OOB signals, D10.2, ALIGNs)

XL (link layer for expander phys) state machine

XL transmitter (AIP, BREAK, BROADCAST, CLOSE, OPEN_ACCEPT, OPEN_REJECT, OPEN address frame, dword, idle dword)

Link reset sequence complete

Phy reset sequence complete

At each level, there is feedback to indicate when a dword has been transmitted and the next dword may be presented.

Key:
- - - - - Controls
──────── Data path

# Expander device model

- Expander function divided into 3 blocks
  - Expander connection router (ECR)
  - Expander connection manager (ECM)
  - Broadcast primitive processor (BPP)



**Expander device**

**Internal expander port**

Virtual phy
- Expander link layer (XL)
- Protocol-specific link and transport layers

Narrow ports or wide ports

**External expander port - SATA host port**

External SATA phy
- STP/SATA transport layer
- Expander link layer (XL and SL_IR)
- SATA phy layer

Narrow ports

**External expander port**

External phy
- Expander link layer (XL and SL_IR)
- Phy layer (SP and DWS)

Narrow ports or wide ports

**Expander function**
- Expander connection router (ECR)
- Expander connection manager (ECM)
- Broadcast primitive processor (BPP)

# Expander device interface details

- XL state machine in each expander phy interfaces to the expander function
- Requests forwarded to peer phys as indications
- Responses forwarded to peer phys as confirmations
- ECR just forwards; ECM and BPP generate own replies
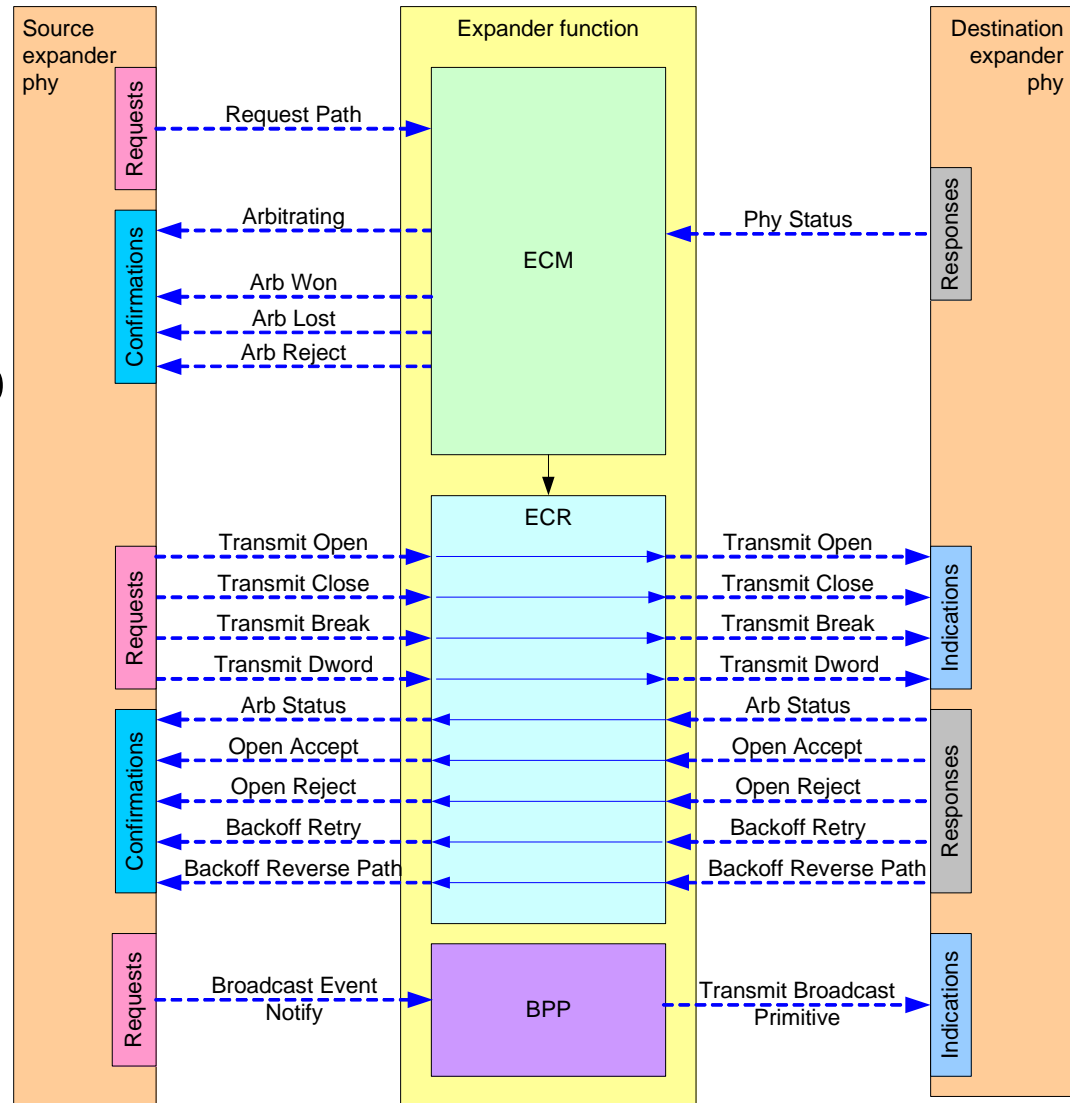
# Expander routing attributes and methods

- Each expander phy has an expander routing attribute
- The attribute determines the routing methods the ECM uses with each phy

| Routing attribute | When attached to | Routing method used |
|---|---|---|
| Direct | End device | Direct |
| Table | Expander device | Table + Direct |
| | End device | Direct |
| Subtractive | Expander device | Subtractive (+ Direct) |
| | End device | Direct |

- Direct = route requests to the attached SAS port through this phy
- Table = route requests that match in routing table through this phy
- Subtractive = route unresolved requests through this phy

# Expander route table contents

**Expander route table**

|  | 0 | 1 | 2 | ... | N |
|---|---|---|---|---|---|
| **0** | | | | | |
| **1** | | | | | |
| **2** | | | | | |
| **...** | | | | | |
| **M** | | | | | |

**Phy identifier**

A phy identifier for each expander phy of the expander device.

$N \leq$ number of phys - 1

**Expander route entry**

Includes:
- Routed SAS address
- Enable/disable bit

**Expander route index**

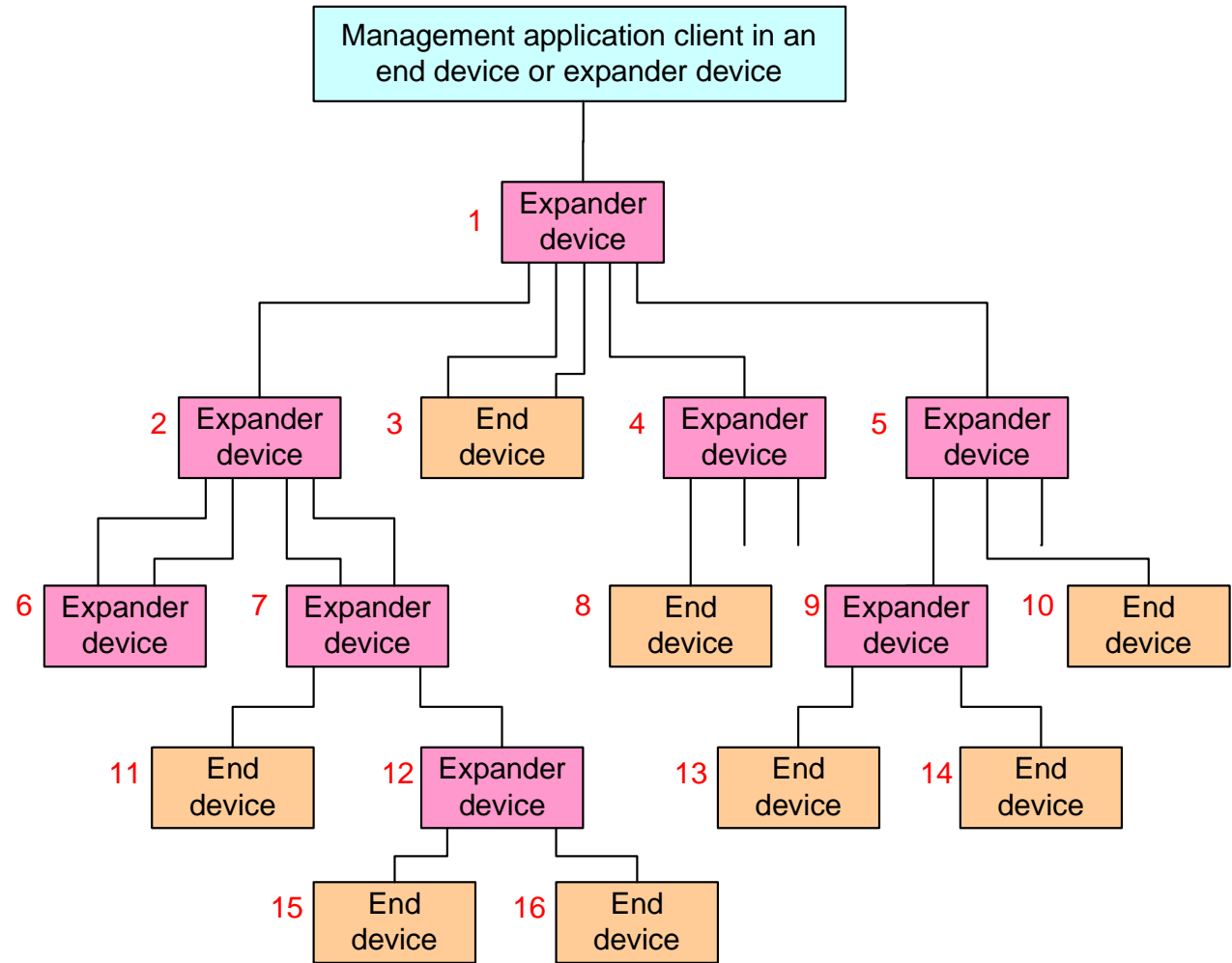An expander route index for each expander route entry.

$M$ = expander route indexes - 1

# Discover process

- Probe the SAS domain one level at a time
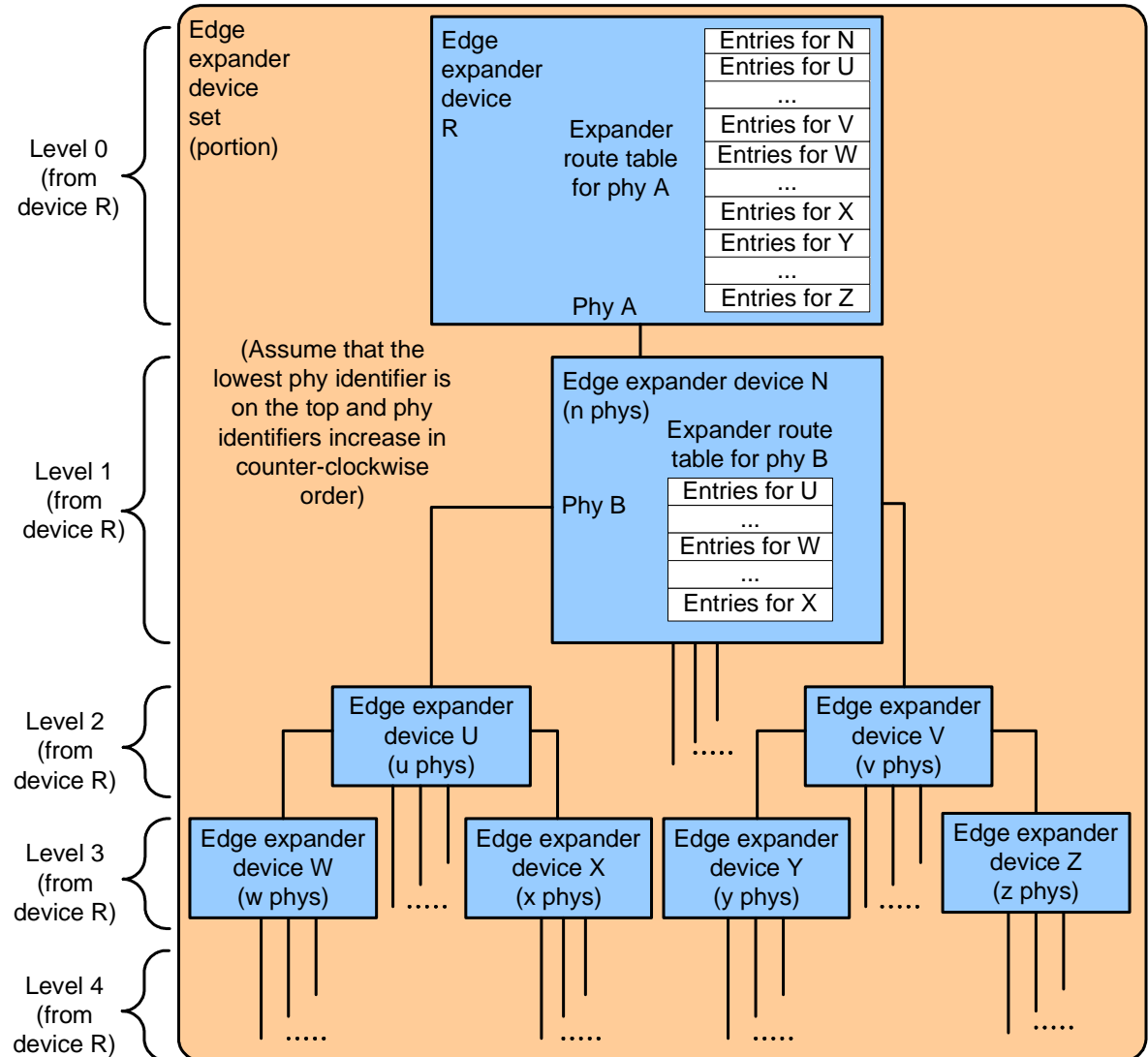- When expander devices with configurable routing tables are found, fill in the tables



Assume that the lowest phy identifier in each expander device is on the top right, and the remaining phys are numbered counter-clockwise

# Expander route table example

- No matter which initiator fills in the table for a phy, it ends up with the same entries in the same locations

# Wrap up

# Additional tutorials

- General overview
- More detailed 7-part tutorial:
  - #1: General introduction and architectural overview
  - #2: Physical and phy layers
  - #3: Link layer
    - Part 1) Primitives, address frames, connections
    - Part 2) Arbitration fairness, deadlocks and livelocks, rate matching, SSP, STP, and SMP frame transmission
  - #4: Upper layers
    - Part 1) SCSI application and transport layers
    - Part 2) ATA application and transport layers
    - Part 3) Management application and transport layers, plus port layer

# Key SCSI standards

- Working drafts of SCSI standards are available on http://www.t10.org
- Published through http://www.incits.org
  - Serial Attached SCSI
  - SCSI Architecture Model - 3
  - SCSI Primary Commands - 3
  - SCSI Block Commands - 2
  - SCSI Stream Commands - 2
  - SCSI Enclosure Services – 2
- SAS connector specifications are available on http://www.sffcommittee.org
  - SFF 8482 (internal)
  - SFF 8470 (external)
  - SFF 8223, 8224, 8225 (form factors)

# Key ATA standards

- Working drafts of ATA standards are available on http://www.t13.org
  - Serial ATA 1.0a (private WG specification)
  - ATA/ATAPI-7 Volume 1 (architecture and commands)
  - ATA/ATAPI-7 Volume 3 (Serial ATA standard)
- Serial ATA II specifications are available on http://www.t10.org and http://www.serialata.org
  - Serial ATA II: Extensions to Serial ATA 1.0
  - Serial ATA II: Port Multiplier
  - Serial ATA II: Port Selector
  - Serial ATA II: Cables and Connectors Volume 1

# For more information

- International Committee for Information Technology Standards
  - http://www.incits.org
- T10 (SCSI standards)
  - http://www.t10.org
  - Latest SAS working draft
  - T10 reflector for developers
- T13 (ATA standards)
  - http://www.t13.org
  - T13 reflector for developers
- T11 (Fibre Channel standards)
  - http://www.t11.org
- SFF (connectors)
  - http://www.sffcommittee.org

- SCSI Trade Association
  - http://www.scsita.org
- Serial ATA Working Group
  - http://www.serialata.org
- SNIA (Storage Networking Industry Association)
  - http://www.snia.org
- Industry news
  - http://www.infostor.com
  - http://www.byteandswitch.com
  - http://www.wwpi.com
  - http://searchstorage.com
- Training
  - http://www.knowledgetek.com