FlexE 技术白皮书

中兴通讯股份有限公司

ZTE CORPORATION

修改记录

文件编号	版本号	拟制人/修改人	拟制日期/修改日期	更改理由	主要更改内容（写要点即可）
	V1.0	刘峰/成剑	2017.08.18	初稿	
注：文件第一次归档时，“更改理由”、“主要更改内容”栏写“无”。					

目录

FLEXE 技术白皮书

摘要： 本文给出灵活以太网的背景知识，标准发展历程、内容，应用领域、组网模型等
关键词：灵活以太网（FlexE）、FlexE tunnel

1 前言

在过去 30 年中，以太网的速度从 10M 提升到 100GE：10M、100M、GE、10GE、40GE、100GE，基本是每 10 年速率 10 增长倍的发展趋势。在最近 3 到 5 年时间里，以太网新速率开始呈现多维度演进，业界开始对另外新出现的 6 种速率的以太网感兴趣：从 2.5GE 到 400GE，包括 2.5GE、5GE、25GE、50GE、200GE、400GE。在 2016 年，业界就引入了 3 种以太网速率（2.5GE、5GE、25GE），目前在进展中的标准还有 3 种以太网速率（50GE、200GE、400GE）。以太网的发展主要有以下三大驱动力：数据中心的应用推动了 25GE- 400GE 以太网的演进；电信运营商网络对于高速率、大带宽的需求推动了 50GE、200GE 和 400GE 以太网的发展；企业的高速率接入侧应用推动了 2.5GE 和 5GE 以太网的发展。此外，还有一些行业应用场景也在催生新的以太网接口，例如车载以太网、工业以太网、以太网供电等方面。

以太网的发展历程大致可以分为以下三个阶段：

第一阶段是原生以太网（Native Ethernet）：从 1980 年产生，广泛应用于园区、企业以及数据中心互联；基于 IEEE 802.3/1 的开放标准，支持互联互通产品；此后进一步延伸到 HPC、存储和垂直应用领域，从而催生出业界最广泛、强大的生态系统。

第二阶段是电信以太网（Carrier Ethernet）：从 2000 年发展至今，面向运营商网络应用提供电信级的城域网、3G/4G/4.5G 承载网和专线接入服务；通过引入 IP/MPLS 技术，具备了 QoS/QoE 保障、OAM、保护倒换和高性能时钟等电信级功能，使运营商网络能够实现低成本、高可靠建网及降低维护成本。

第三阶段是灵活以太网（Flexible Ethernet）：从 2015 年起步，面向 5G 网络中的云服务、网络切片、AR/VR/超高清视频等时延敏感业务需求；通过接口技术创新，实现高速大端口 400GE、1TE 等演进以及通道化实现子速率承载、硬管道及隔离以及时延敏感网络技术；进一步构建智能端到端链路，实现可保障的 IP 低时延、高 QoS 服务的数据网络。

以太网先天具备的实现灵活、操作简化和部署经济等优势，在全球获得广泛普及并在不断完善生态。面向未来移动万物互联的发展需求，驱动带宽速率大幅提升。作为面向未来网络的核心承载技术，以太网技术将加速创新和演进，为面向未来全 IP 网络承载奠定基石。IP 网最重要的发展趋势是以太网化，数据业务的以太网化已经成为全球主导趋势；以太网技术的加速创新和演进，为面向未来全 IP 网络承载奠定了基石。面向未来，灵活以太网将成为未来网络发展的关键方向，基于分片技术的灵活以太网成为未来趋势。

以太网接口的网络分片解决方案在大型网络中的应用，将实现带宽弹性、灵活分配（分片及绑定）；专用硬管道，保证服务质量和安全；提供低延时解决方案；能够与 SDN 技术融合，实现网络动态调整。运营商网络中，基于接口通道化技术的灵活以太网可以提供网络分片、子接口隔离等功能，支持基于业务体验的未来网络架构，是未来网络演进、发展的主要方向。

2011 年 1 月成立灵活以太网研究小组，2015 年 7 月发布草案，2016 年 3 月发布灵活以太网的 1.0 标准内容（**OIF-FLEXE-01.0**），目前正在起草 2.0 标准内容。

2 技术介绍

2.1 技术背景

过去以太网技术标准中，以太网 MAC 报文速率和 PHY 的速率始终保持配合和一致，两者速率同步发展。但是当以太网业务速率提升到 100GE 以上时遇到新问题：物理 PHY 的速度发展遇到瓶颈，速度提升缓慢下来，而且 PHY 的价格下降缓慢，高速 PHY 的性价比没有提高，反而在降低。例如，400GE 速率的光模块价格超过了 4 个 100GE 光模块的价格，导致 400GE 光模块商用时的性价比降低了，在经济上不如使用 4 个 100GE 的光模块。

FlexE 技术实现业务速率和物理通道速率的解耦，物理接口速率不再等于客户业务速率，而是其他速率（比如客户业务速率是 400GE，但物理通道 PHY 的速率是 100GE 或其他速率），物理接口速率可以是灵活的，比如 n*100G 或 n*200G。客户业务不一定在一个物理通道上传递，而是由多个物理通道捆绑起来形成一个虚拟的逻辑通道来传递。业务速率和物理通道速率解耦后，客户业务速度可以是多样的，物理通道的速率也是多种速率，相互独立，这样大带宽的客户业务可以由多个低速物理通道捆绑起来进行传递，解决了高速物理通道性价比不高的问题，如图 2-1。

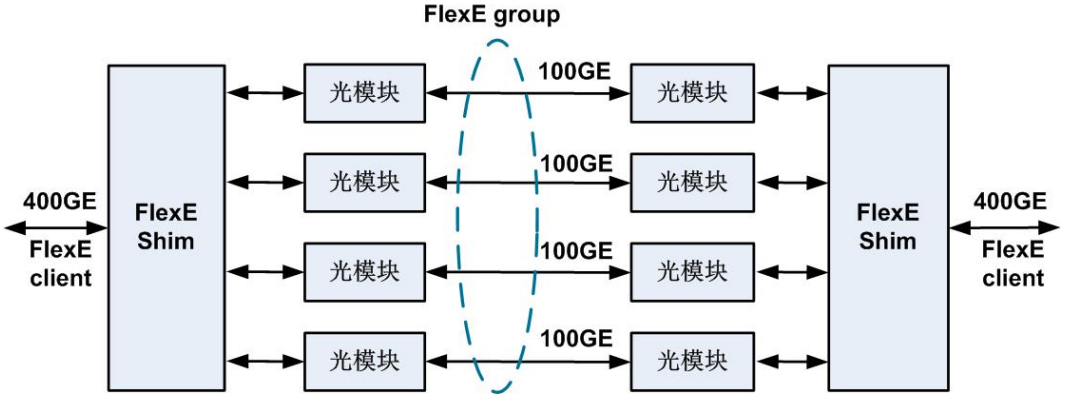


图 2-1

采用 FlexE 技术，除了实现大速率的客户业务通过多个低速物理通道进行传递，也可以实现多条低速率客户业务在一个物理通道共享传递，多条业务流之间是分片隔离，互不影响，实现网络分片功能，提高了业务传送效率和物理隔离需求。

FlexE 技术是在 IEEE802.3 的协议栈的 RS 和 PCS 层增加一个 FlexE Shim 层，将业务逻辑层和物理层隔开，通过绑定多条 100G PHY 来传输大流量的以太网业务，这样逻辑层面可以实现大业务速率、子速率、通道化等功能，以及网络切片需求。

2.2 FlexE 技术层次位置

在 100GE 以太网业务流传递中，以太网数据报文（MAC）业务流经过 RS 层连接 PHY，在 PHY 层经过 PSC、PMA、PMA 功能模块后发送出去，如图 2-2 中左边部分，其中在 PCS 功能模块中，对业务流进行 64/66 编码，然后是扰码，lane 分配和 AM 信息块的插入。FlexE 技术是在原以太网业务处理流程的 MAC 层和 PCS 层之间增加了 FlexE shim 层，如图 2-2 中右边部分。FlexE shim 层由 64/66 编码、时隙排列、成员分发和开销插入四个部分组成，其中 FlexE shim 层的 64/66 编码和 PCS 的 64/66 编码是相同功能，因此在 FlexE shim 层中实现了 64/66 编码功能后，PCS 中的 64/66 编码可以省去。

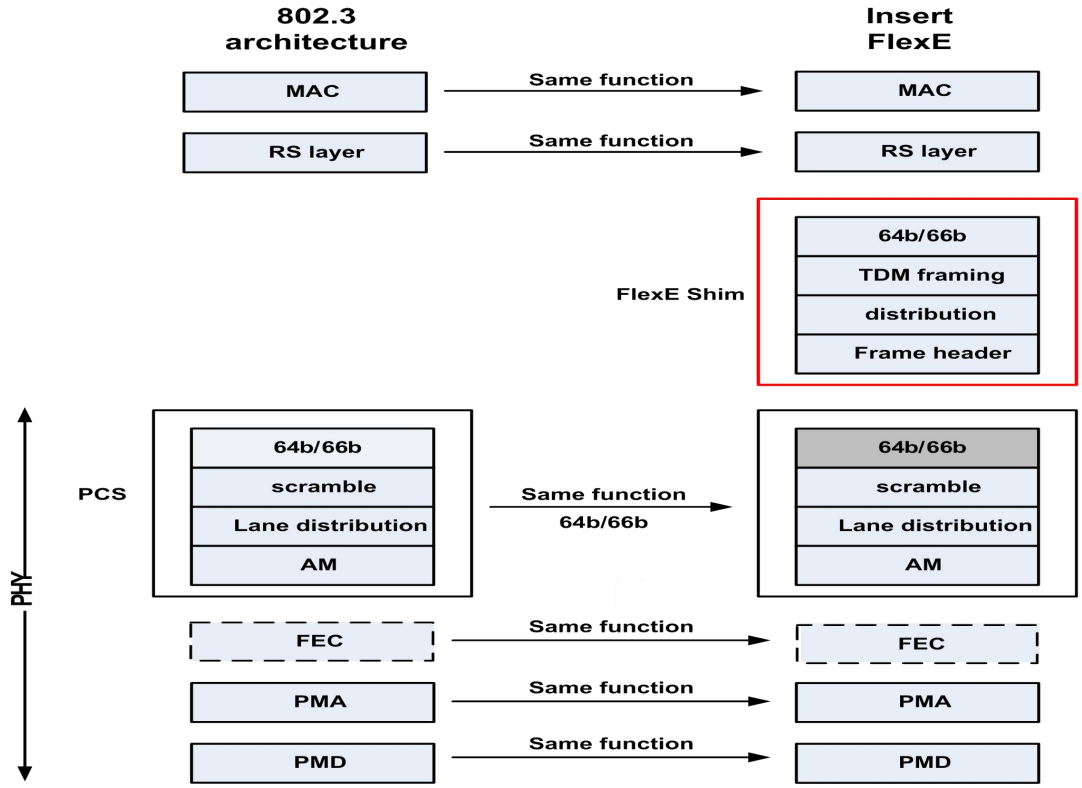


图 2-2

2.3 FlexE 技术应用方式

应用 FlexE 技术可以实现三种应用模式：链路捆绑、子速率和通道化。在一个应用事例中可以只涉及三种应用模式的一种，也可以涉及多种应用模式。

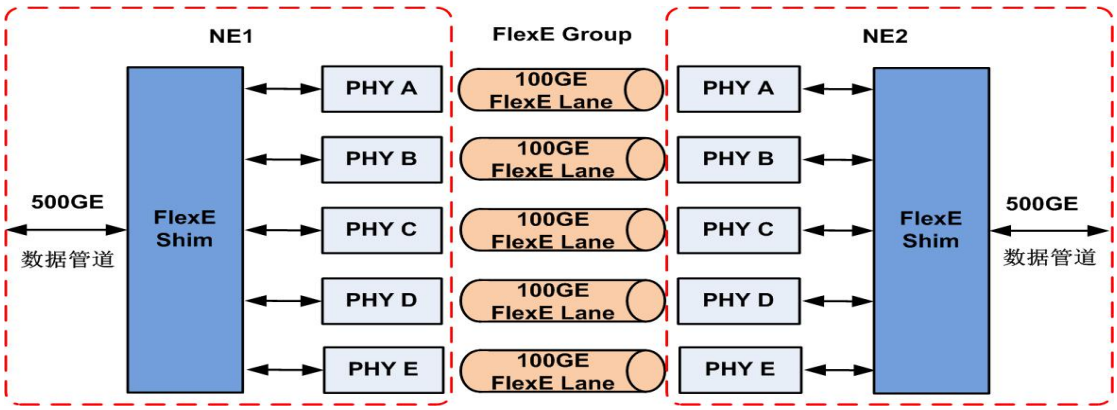


图 2-3

链路捆绑是将多个物理通道捆绑起来，形成一个大的逻辑通道，实现大流量的业务传输，如图 2-3。链路捆绑模式类似于 LAG 协议，LAG 协议是将以太网报文通过 HASH 等算法将报文分配到不同物理管道上传输，LAG 协议是在报文层面分配报文到不同物理通道，依赖报文 MAC 地址等信息和 HASH 算法，存在分配不均匀，甚至无法分配的现象（当业务报文的 MAC 等信息固定时），传输效率低、延迟时间较长。FlexE 技术的链路捆绑模式，是将 66 比特块分配到不同物理通道，按照时隙模式进行分配，业务分配是严格均匀的，传输效率高，延迟时间短。FlexE 技术的链路捆绑模式实现多个低速率物理管道来传递高速客户业务。

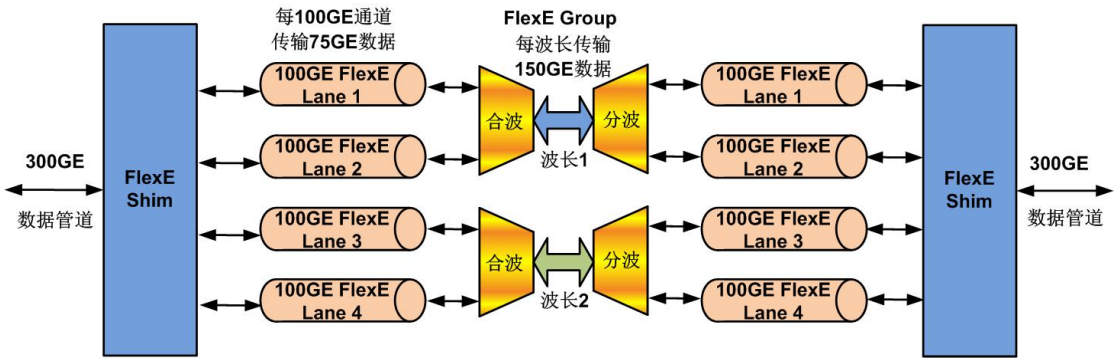


图 2-4

子速率模式是单条客户业务速率小于一条物理通道速率，将多条客户速率汇聚起来共享一条物理通道，提高物理通道的带宽利用率，如图 2-4。多条客户流共享一条物理通道，在物理通道的不同时隙上分别传递多个客户业务，多条客户业务流采用不同时隙，实现业务隔离，等效于物理隔离。子速率模式提高了物理通道的传递效率，实现网络切片功能。

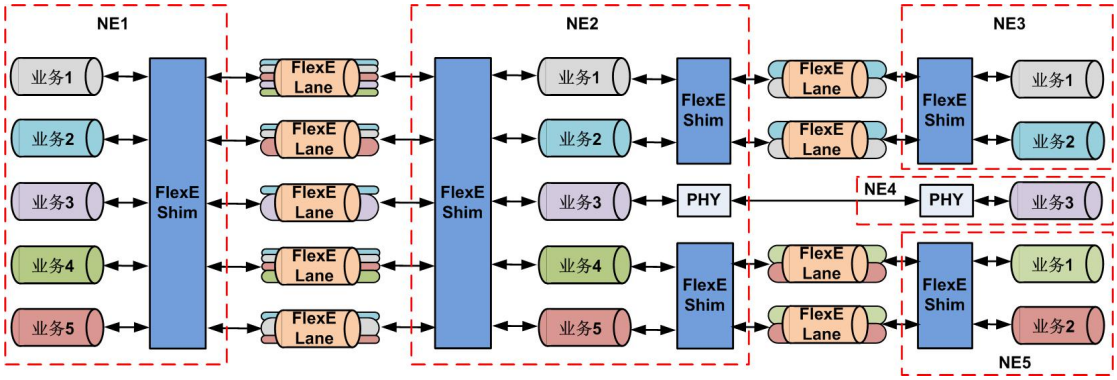


图 2-5

通道化模式是客户业务在多条物理通道上的多个时隙传递，客户业务分布在多条不同物理通道的多条时隙上，多个客户共享多条物理通道，如图 2-5。客户在 FlexE 上传递时，根据实际情况选择不同的时隙组合，合理利用物理通道带宽。

2.4 Shim 层

FlexE 协议定义了一个时分复用的 FlexE Shim 层，FlexE Shim 通过多个绑定的 PHY 来承载各种 IEEE 定义的以太网业务（FlexE Client），Shim 层可以支持各种的以太网 MAC 报文，包括大于或小于单个物理 PHY 速率的以太网报文。客户业务承载过程如图 2-4，以太网报文进行 64/66 编码，然后通过 idle 块的插入和删除进行速度适配，将业务插入到 master calendar 中。master calendar 将所有时隙分配成多个 sub calendar（成员），添加 FlexE 开销，扰码后经过 PMA、PMD 发送出去。

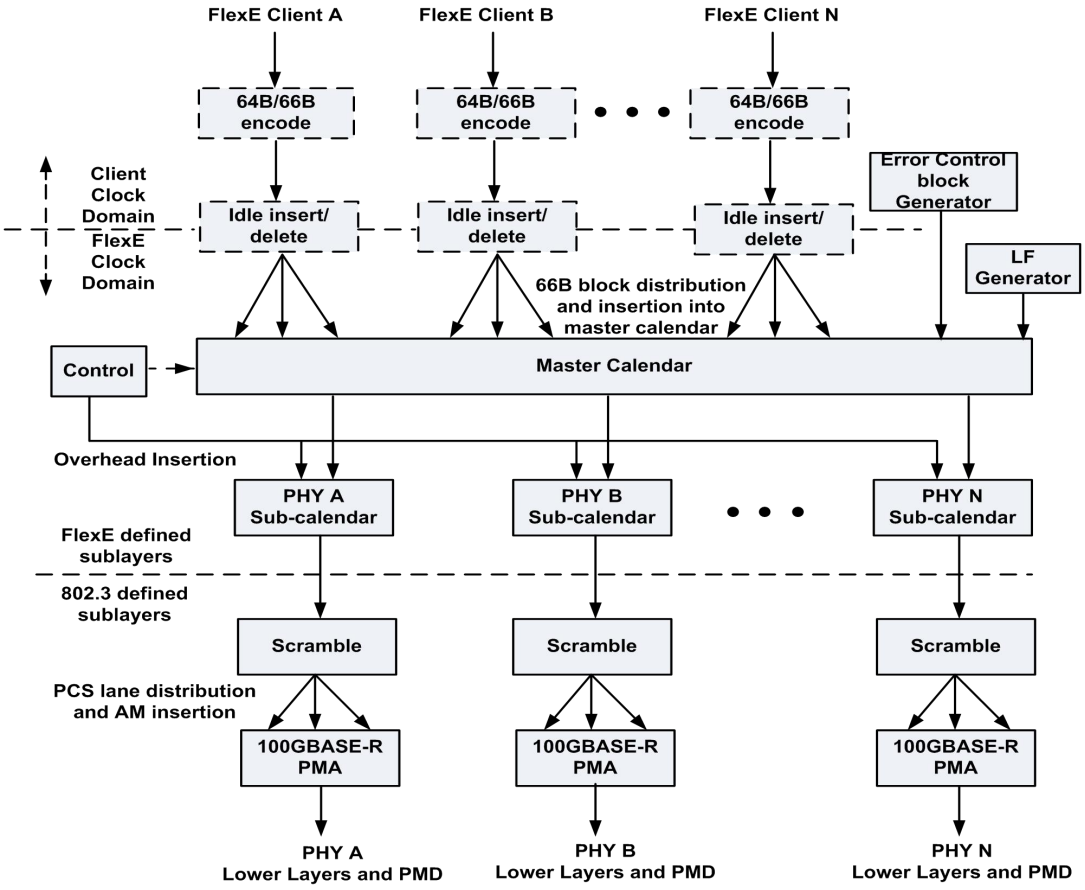


图 2-4

在接收端，如图 2-5，从 PMD、PMA 上恢复信号，经过解扰码，恢复出 66 比特块，寻找 FlexE 开销块，确定 sub calendar，所有 sub calendar 拼装出 master calendar，从中找出每条客户业务流，然后通过 idle 块的插入和删除进行速度调整，进行 64/66 反编码，恢复出原始客户业务。

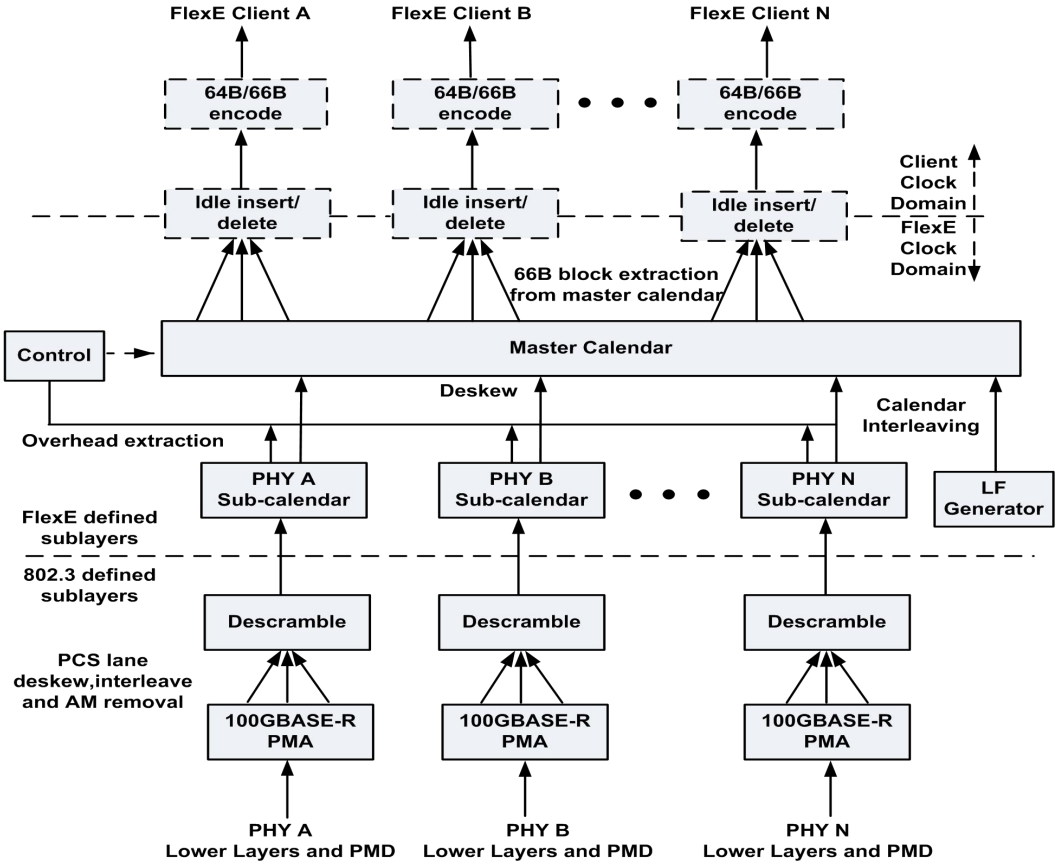


图 2-5

2.5 Calendar 结构

FlexE 扩展子层的 calendar 包括两个部分： sub calendar 和 master calendar。

FlexE 协议定义每个物理成员 PHY 上传递一个 sub calendar，sub calendar 按照 20 个 5GE 时隙来划分（注：FlexE V1.0 标准定义物理 PHY 速率是 100GE,且不支持不同速率 PHY 的混合应用）。FlexE Shim 层是一个 master calendar，由多个 sub calendar 组成，master calendar 有 n*20 个 5GE 时隙(注:n 为捆绑组的总成员数)。FlexE client 的 64b/66b 按照时隙方式间插到 master calendar 中。（注：FlexE V1.0 标准定义 FlexE Client 对应的速率为标准 MAC 速率：10G，25G，40G，n*50G 等，实际应用 FlexE client 速率可支持基于 5G 时隙颗粒度的整数倍速率）

FlexE sub_calendar 结构如图 2-6， 每个 PHY 上有 20 个时隙，每个时隙是一个 66 比特的数据块组成的 TDM 码块流，速率为 5Gbps。每间隔 1023×20 个 66 比特块插入一个 FlexE 开销块：Overhead。FlexE 开销块是一个 66 比特的信息块，开销块用于定位时隙位置，以及不同成员之间的时隙对齐。

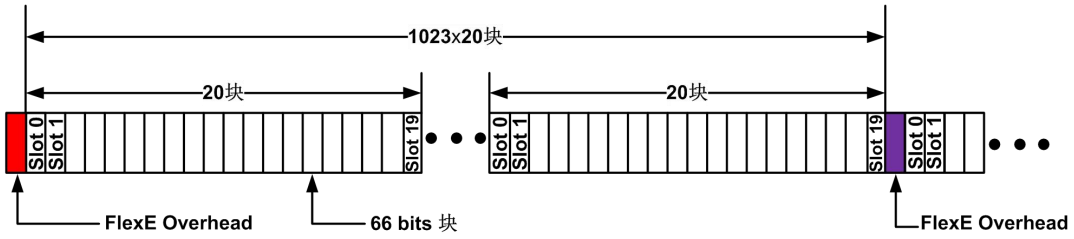


图 2-6

Master calendar 的结构定义如图 2-7。Master calendar 标识该 FlexE group 所包含的所有 sub calendar 的时隙之和，当 FlexE group 有 n 个成员时，master calendar 有 n*20 个时隙。它是对应 FlexE client 的资源池，FlexE client 可指定该 calendar 中任何时隙组合带宽来承载。

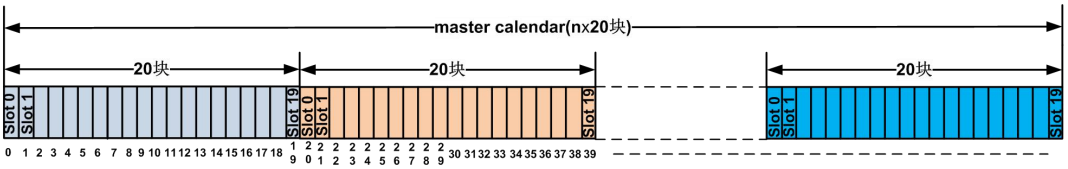


图 2-7

Master calendar 与 sub calendar 的时隙分发关系如图 2-8。如图例 master calendar 包含 4 个 PHY 共 80 个时隙，master calendar 以连续 20 个时隙为一组，分配到 4 个 sub calendar 承载。每个成员的开销字节中携带有本成员的编号 PHY number，成员之间按照 PHY number 编号从小到大的次序进行排序（注：PHY number 编号不要求是连续）。

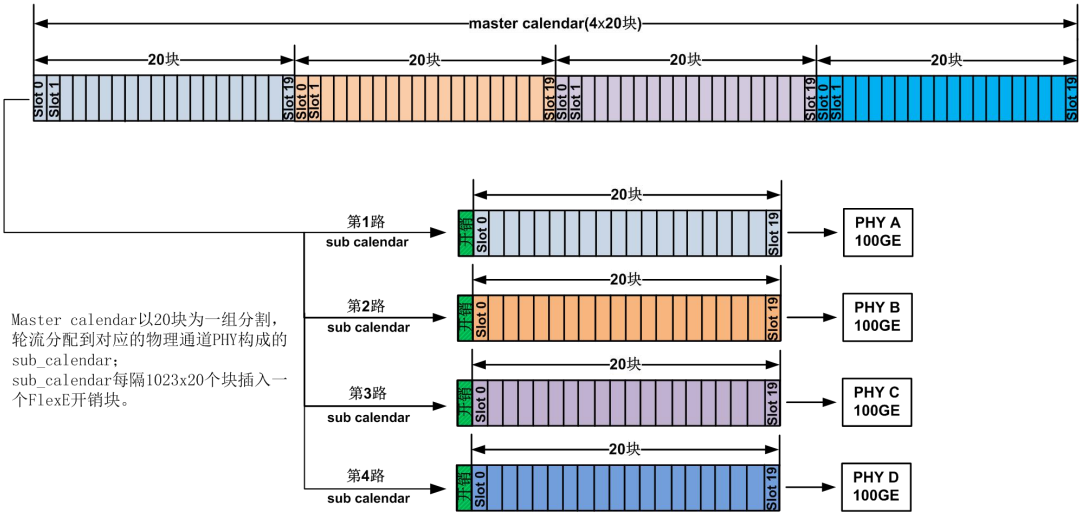


图 2-8

2.6 开销结构

FlexE 开销块是一个 66 比特的信息块，每间隔 1023×20 个 66 比特块有一个开销块，连续 8 个开销块组成一个 FlexE 帧，连续 32 个 FlexE 帧组成一个 FlexE 复帧，如图 2-9。

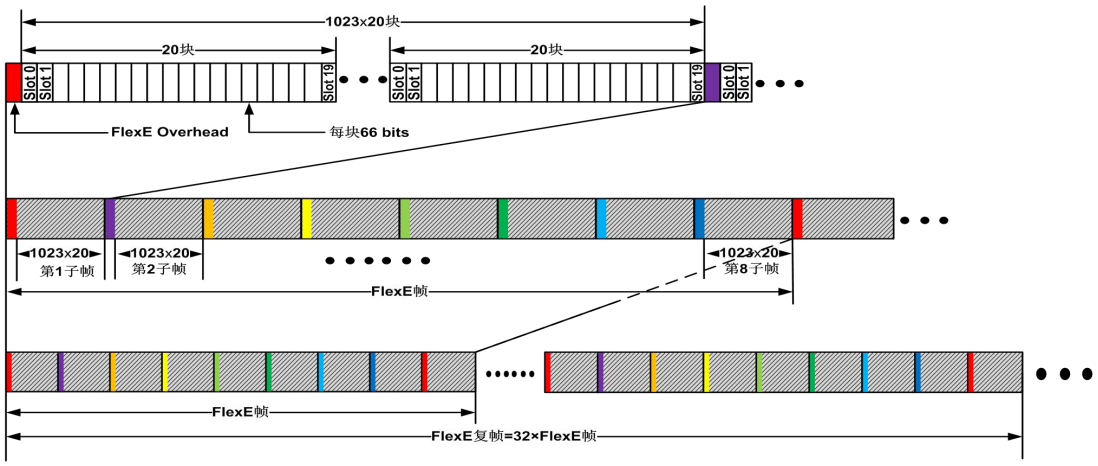


图 2-9

在 FlexE 开销结构中，由第一个开销块来指示帧头，第一个开销块是通过 0x4B 和 0x5 两个部分来指示。当检测到 66 比特块以 10 开始，对应位置的内容是 0x4B 和 0x5，则确定该信息块是一个 FlexE 的开销帧的首块，后续的 7 个开销块则是帧结构中第 2-8 个开销块。如图 2-10。

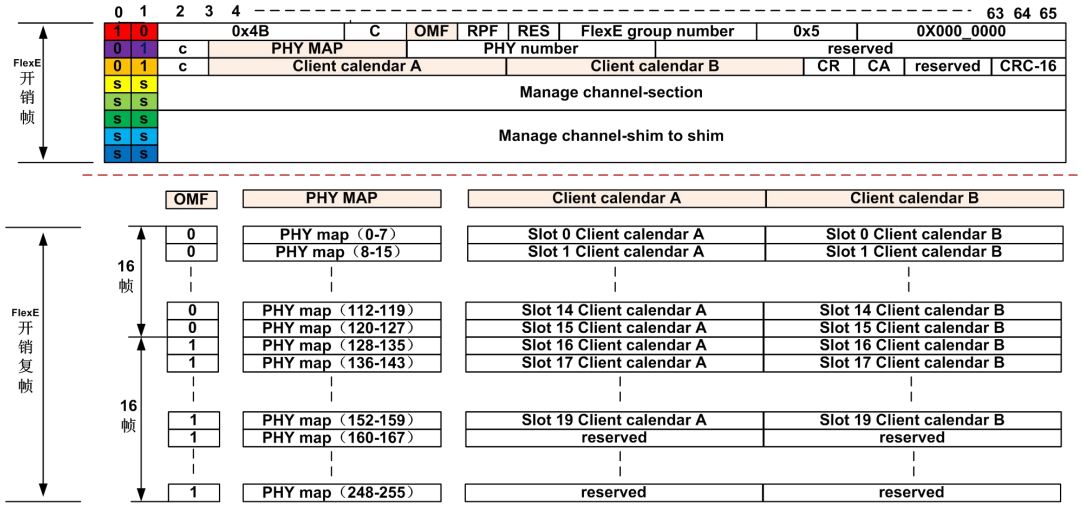


图 2-10

FlexE 开销结构中携带成员组编号、成员编号、时隙属性等各类信息，具体字段定义如下：

- C：3 比特，位置在第一个 66b 开销块的第 10 个比特，以及第二第三 66b 开销块的第 3 个比特，分开存放的原因主要在于防止突发错误的干扰，其作用是用来表征目前配置的 Calendar 类型，0 为 A Calendar configuration, 1 为 B Calendar configuration，使用三个比特是用多数判决的思想来判定 Calendar 类型。
- OMF：1 比特，位置在第一个 66b 开销块的第 11 个比特，开销复帧指示位，对于 FlexE 的有些开销是分布在连续的 32 个 FlexE 的开销中的（用来表征 256 个 PHY，位于 PHY MAP 中，每个 8 比特，一共需要 32 个），被称为 FlexE 开销复帧，OMF 是用来指示这些 FlexE 开销的，前 16 个 FlexE 开销中的 OMF 为 0，后 16 个 FlexE 开销中的 OMF 为 1。
- FlexE group number：20 比特，位置在第一个 66b 开销块的第 14 比特到第

33 比特，用来区别存在多个 FlexE group 的情况，因为对于每个 FlexE group 来说，里面的 PHY 的数量都可以设为 1-255 之间的数字，如果没有区分，有可能接收到两个 PHY，并且他们的编号一致，导致无法区分。当使用这部分开销时，收发两端必须协商好，使两端的 FlexE group number 保持一致。

- PHY MAP：在一个开销帧中，PHY MAP 有 8 个比特，通过复帧结构（32 帧组成一个复帧），PHY MAP 在一个复帧中共 256 比特（FlexE group 最多可以绑定 256 个 PHY，0 和 255 作为特殊用途，可用的为 1-254，实际中可能会用到 4-8 条 PHY），位置在第二个 66b 开销块的第 3 比特到第 10 比特，一共分 32 次传送完毕，256 比特的每个比特表征对应 PHY 的使用情况，为 1 表示使用，0 表示未使用，对于使用的 PHY 的编号并不需要连续。
- PHY number：8 比特，位置在第二个 66b 开销块的第 11 比特到第 18 比特，用来表示本物理成员 PHY 的编号，编号在一个 group 中必须是唯一的，该部分开销需要两次确认。
- Client Calendar A 和 Client Calendar B：每个开销块携带 16 比特的 A 和 16 比特的 B，通过复帧结构（32 帧组成一个复帧），分 32 次传送完毕，用来表征 slot 里面所装载的客户业务类型，因为对于每个 PHY 有 20 个 slots，每个 slot 装的业务可能装不同类型的业务，在收端需要将属于同一个客户业务的 slot 收集起来恢复客户业务，用 Client Calendar 来表征该 slot 所装的业务是哪种客户业务。由于每个 PHY 有 20 个 slot，需要 20 个开销帧进行传送，32 个复帧中前 20 帧传递 Client Calendar，剩下的 12 个开销帧作为保留。0x0000 表示该 slot 是 unused，但还可以使用，0xFFFF 表示该 slot 是 unavailable 的，不能再被使用。至于具体业务类型对应的码字，标准未做规定。
- CR,CA：calendar request 和 calendar acknowledge 用作增加或者删除 slot 中客户业务的请求和反馈信息。
- Manage channel：5 个 66b 开销块。分为两个部分，一部分为段层的管理通道，占用 2 个 66b 开销块，位于第四和第五个开销块中，带宽为 1.26Mb/s，另一部分为 shim 到 shim 的管理通道，占用 3 个 66b 开销块，位于第六到第八个开销块中，带宽为 1.89Mb/s。对于管理通道里面的内容，标准并不做规定，每个 PHY 都有自己的管理通道，这些管理通道不会进行聚合。对于 FlexE aware 场景，OTN 设备入口处，会终结段层的管理通道，提取出相关的信息，并将段层管理通道的位置填上空闲控制块，在 OTN 设备的出口处，会重新在段层管理通道上重新插入开销。
- RPF：1 比特，位于第一个 66b 块的 12 比特的位置，远端 PHY 故障指示，用来向远端 shim 通知在本地检测到 PHY 失效。
- CRC-16：开销帧前三行的内容进行 CRC-16 的运算结果，用于检测前三行内容是否有误码。

2.7 时隙帧结构

FlexE 协议定义每个物理成员 PHY(注：标准为 100GE)上传递一个 sub calendar，sub calendar 按照 0、1、2 ... ...18、19，0、1、2 ... ... 18、19 重复 20 个编号规则来划分 66b 码块顺序，同一编号的码块在逻辑上组成一个独立的时隙，在 FlexE shim 中作为一个独立物理带宽资源单元分配使用。按标准 100GE 的 PHY 计算，按 20 个单元划分，每个独立单元时隙带宽为 5Gbps。Sub calendar 每间隔 1023×20 个 66b 码块插入一个 FlexE 开销块 Overhead，FlexE 开销块用于定位时隙位置，以及不同成员之间的时隙对齐。

FlexE sub calendar 的时隙帧结构如图 2-11。

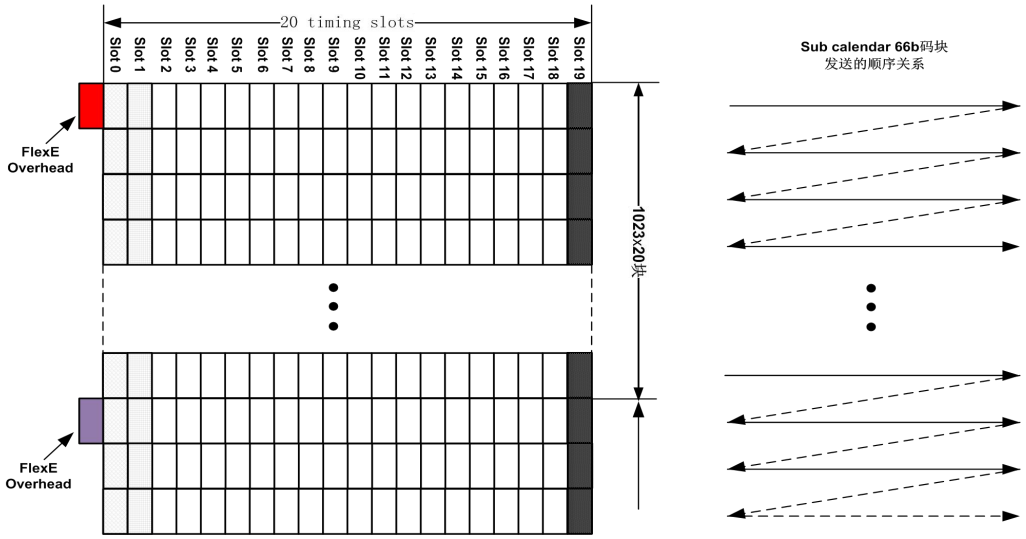


图 2-11

2.8 管理通道

在 FlexE 开销帧中包含两个管理通道字段，分别是 Section 层的管理通道和 Shim-to-Shim 的管理通道，见图 2-10。Section 层的管理通道字节位于第 4 行和第 5 行，共 16 个字节，Section 层的管理通道带宽是 1.260 Mb/s。Section 层的管理通道信息在相邻度的段层处终结，只在相邻的物理管道之间有效，不会穿透第三者网络。Shim-to-Shim 的管理通道字节位于第 6 行、第 7 行、第 8 行，共 24 个字节，Shim-to-Shim 的管理通道带宽是 1.890 Mb/s。Shim-to-Shim 的管理通道信息在相邻的两个 shim 之间有效，在 Aware 场景下，Shim-to-Shim 管理通道信息会穿透第三者网络。

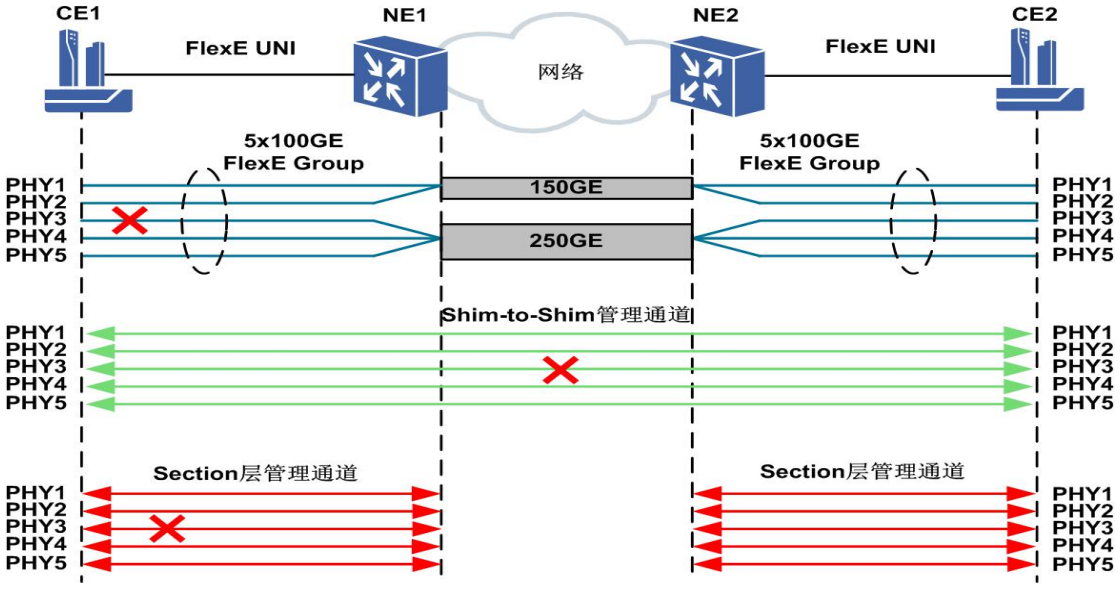


图 2-10

2.9 客户业务调整机制

FlexE 协议中定了 Client Calendar A、Client Calendar B、CR、CA、C 信息，用于实现客户业务的动态增加和删除功能。C 比特有 3 个比特，用多数判断原则决定 C 比特代表“0”还是代表“1”。 当 C 比特代表“0”，表示 Client Calendar A 的配置信息处于工作状态（on line），Client Calendar B 的配置信息是备用状态（off line）；当 C 比特代表“1”，

表示 Client Calendar B 的配置信息处于工作状态，Client Calendar A 的配置信息是备用状态；CR 和 CA 信息用于两端进行配置信息调整时的握手协商指示，CR 表示调整申请，CA 表示允许调整应答。

若有需要进行业务时隙增删、调整时，首先会在处于备用状态的 Calendar 中进行配置，主要是将某些 slot 的 client calendar 值进行重新配置：增加业务的话，部分标记为 0x0000 的 unused slots 会被使用，其 client calendar 开销会被配置为与该业务类型一致的编码；删除时，则将要删除业务占用的 slots 的 client calendar 开销变为 0x0000（unused）。

当 FlexE mux 端 offline Calendar 完成配置操作之后，在进行 Calendar 转换之前，FlexE mux 端和 FlexE demux 端需要进行确认。首先将所有 PHY 中 CR 比特变为 offline Calendar 的类型，从 FlexE mux 端传到 FlexE demux 端，当接收端检测到 CR 字节的变化，知道 FlexE mux 端会进行业务调整操作，当 FlexE demux 端完成准备之后，则将更改所有 PHY 中的 CA 比特以与收到的 CR 比特保持一致，FlexE mux 端检测到 CA 比特变化之后，就知道 FlexE demux 端已经做好准备，便将 offline Calendar 转换为 online Calendar，完成整个操作过程。

3 FlexE 扩展技术方案

3.1 技术扩展需求

传统分组设备传输客户业务采用逐跳转发策略，客户业务在网络中每台设备上都需要接收并存储完整的业务报文，然后根据报文头的路由信息查表转发到下一个节点，这种存储转发方式存在延迟大、抖动不可控问题。当时延敏感业务经过多跳存储转发后，到达目的节点的时延变的不可预知，无法满足 5G 承载时延敏感业务对承载传送的要求。

FlexE 技术标准由 OIF 组织制定，用于数据中心设备之间的互联互通，解决物理链路带宽速率不足的问题，标准制定重点考虑点到点的应用场景需求，在组网应用、端到端承载、业务保护上缺少考虑，因此 FlexE 技术在承载网络中进行组网应用时，技术方案需要进行扩展和完善。

承载网络应用 FlexE 技术传输客户业务模型如图 3-1。客户业务在业务接入节点 PE 节点上网络，根据客户业务的 IP 地址、MAC 地址、端口号等信息实现三层路由和二层交换，选择承载业务的网络路径和物理端口，FlexE 端口分配带宽进行传输。在网络中间节点 P 节点，根据业务传输路径在 L1 层进行交叉连接，实现超低时延传输。在目的节点 PE 点业务落地，根据报文的路由信息选择输出端口。客户业务在承载网络中进行端到端传输时，涉及端到端管道的监控、PCS 层的业务交叉、业务保护等扩展技术。

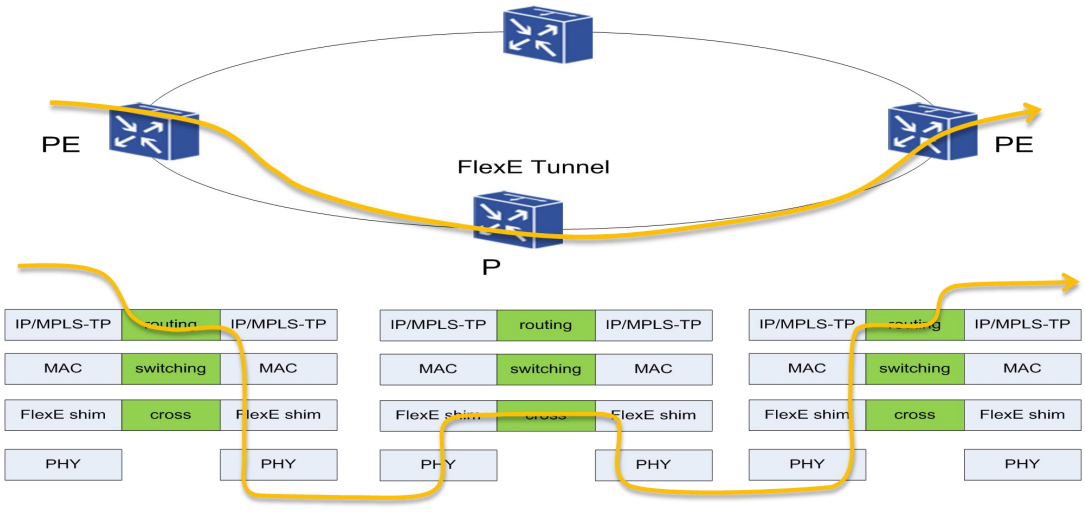


图 3-1

3.2 网络模型

FlexE 技术是一种物理接口的技术，设备之间的物理连接方法，是点到点连接方法，承载网应用 FlexE 技术实现组网，实现客户业务的端到端传输，组网承载模型如图 3-2。在模型中，点到点的 FlexE 技术物理管道连接起来组成一条条客户业务管道，实现客户业务的端到端传输。在 PE 节点，客户业务通过查表路由信息，确定端到端的 FlexE tunnel 传输路径，通过 FlexE group 物理端口传送到下一节点。在 P 点，从上一节点过来的 FlexE group 终结，提取出 FlexE client 业务流，不上送 MAC 层恢复业务报文信息，而是根据配置信息直接进入新的 FlexE group 管道继续进行传送。在目的节点 PE，从 FlexE tunnel 中提取客户业务，根据报文路由信息选择物理端口输出。

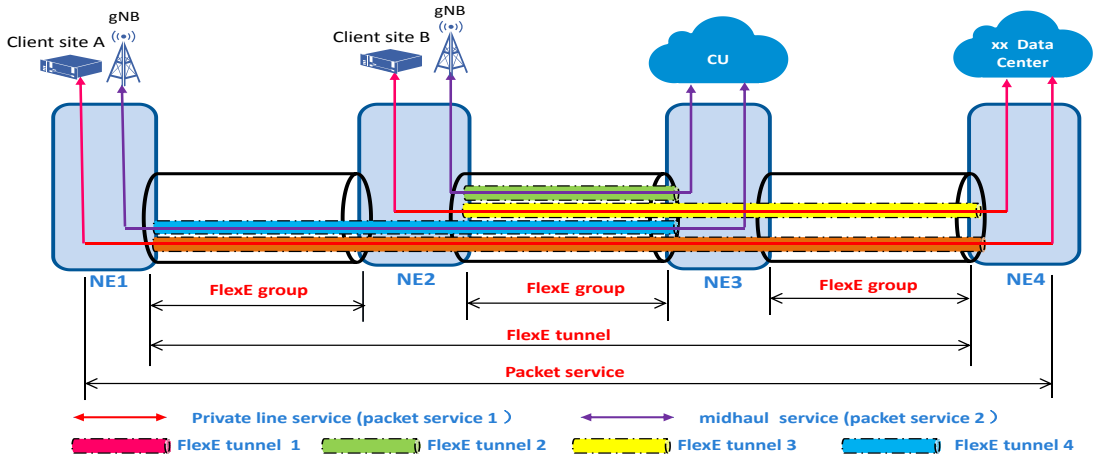


图 3-2

采用 FlexE tunnel 端到端传送技术的 L1 层承载网络模型如图 3-3，FlexE 扩展技术涉及两层网络，FlexE 通道层和 FlexE 段层。

- FlexE 通道层

FlexE 通道层位于 FlexE 客户数据和 FlexE 段层之间，在 FlexE 通道层中，实现客户数据的接入/恢复、增加/删除 OAM 信息、数据流的交叉连接，以及通道的保护。

- FlexE 段层

FlexE 段层位于 FlexE 通道层和 PHY 之间，在 FlexE 段层中，实现接入数据流的速度适配、数据流在 FlexE shim 上映射与解映射、FlexE 帧开销的插入与提取。

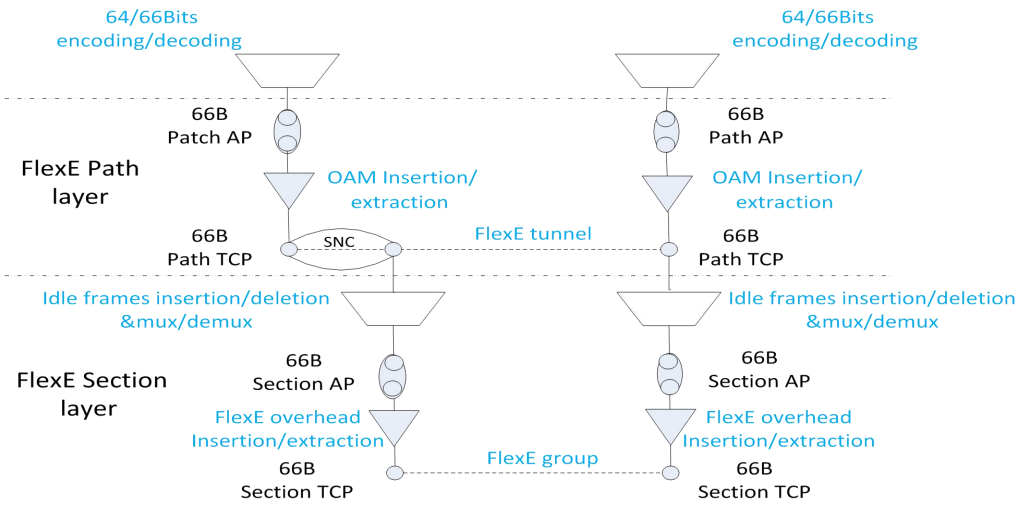


图 3-3

FlexE 通道层形成的端到端连接管道称之为 FlexE tunnel，它是 FlexE client 在 FlexE 层网络中传递的一条逻辑承载管道，客户业务从源节点映射到 FlexE shim，到目的节点从 shim 中解映射之间经过的一条逻辑承载管道。

3.3 技术方案

3.3.1 FlexE client 交叉

端到端 FlexE tunnel 传送技术是构成 L1 层承载网络的基础。端到端 FlexE tunnel 实现的核心是 FlexE client 能从一个 FlexE group 直接在 L1 层交叉到另外一个 FlexE group 承载，而不是上送 MAC 定帧后进行分组交换。

FlexE client 的交叉实现方案如图 3-4。左边 FlexE group 1 表示接收端口，右边 FlexE group 2 表示发送端口。绿色业务 client 在该节点为穿通的 P 节点，配置 client 交叉；绿色业务码流从 FlexE group 1 的 PHY 送到对应的 FlexE master calendar，按配置规则提取对应业务码块流后根据系统交叉配置送到 L1 cross matrix 模块，直接输出映射到 FlexE group 2 的 master calendar 对应的时隙通道，通过 PHY 转发往下一个节点。

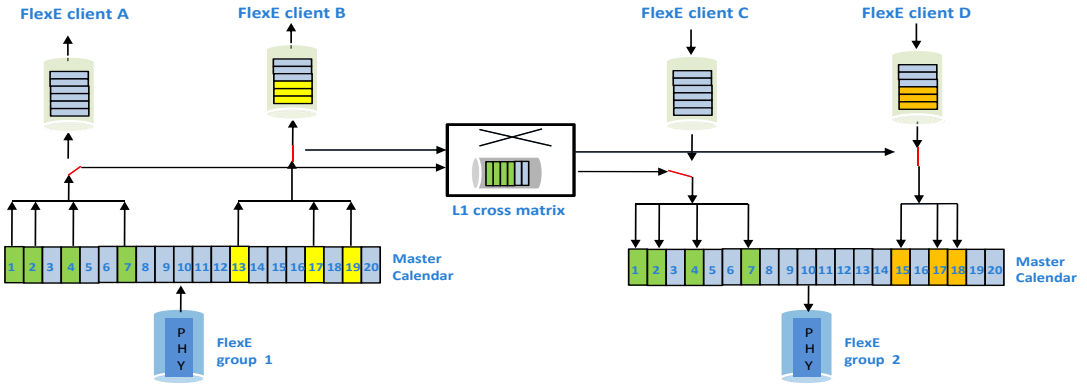


图 3-4

标准 FlexE 技术终结所有 FlexE client，业务上送 MAC 进行分组查表转发；当 FlexE 扩展支持基于 FlexE client 交叉后，在承载网络中的 P 节点，客户业务直接在 L1 层进行交叉，而不再上送 MAC 层，不需要恢复出完整的业务报文。如此，后续的分组设备支持两个转发平面，分组转发和 L1 交叉转发，如图 3-5 所示。

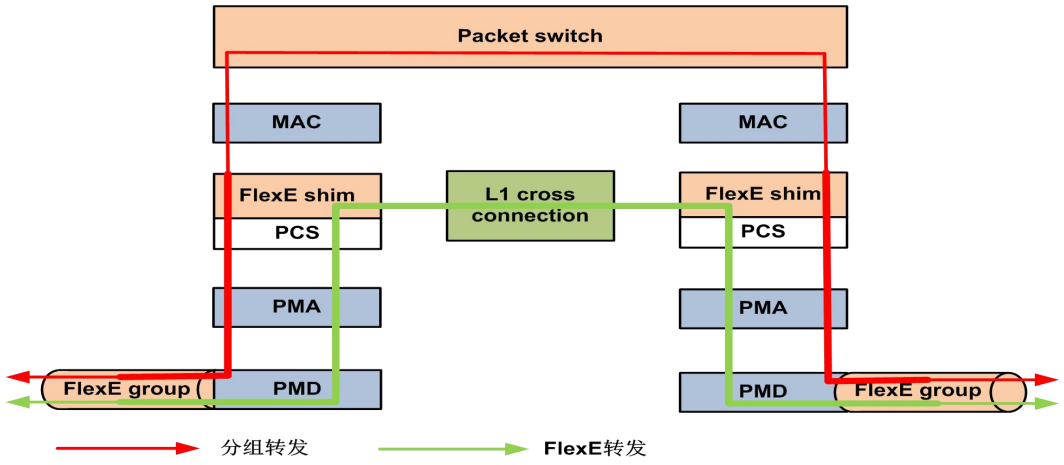


图 3-5

3.3.2 隧道 OAM 功能

采用 FlexE tunnel 隧道技术的承载网络实现客户业务的端到端传输，需实现增加 OAM 开销实现传输管道的端到端监控。在承载网络分层结构中，FlexE 技术涉及两层网络：FlexE 通道层和 FlexE 段层。

FlexE 段层的 OAM 信息来自标准定义的 FlexE 开销帧的内容，通过开销帧头、复帧帧头信息可以提供等效 CC/CV 检测，开销帧中 RPF 信息可以提供远端成员缺陷指示 RDI，通过 PHY map、client calendar A/B 等字段来交互链路带宽以及相关时隙配置业务类型等。详见 2.6 章节定义。

FlexE 通道层 OAM 信息需要进行扩展实现，在客户业务复用进入 FlexE shim 层前，在客户业务流（66b 码块组成的 TDM 码流）中按某种固定周期插入 OAM 信息块，如图 3-7 中标注为 C 的红色码块即为 OAM 块。

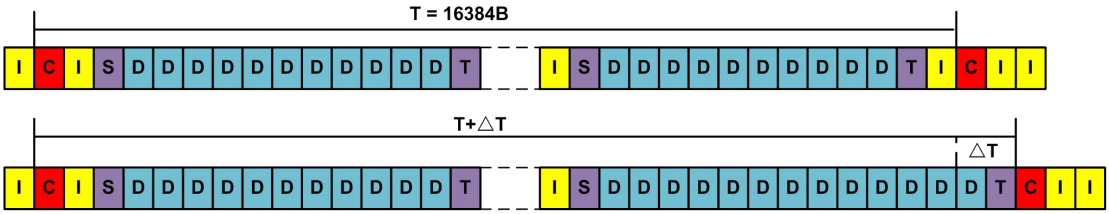


图 3-7

通道层 OAM 是一个特殊的信息块，符合 64/66 编码规范，并且具备特殊的图案标志，可以在接收端被识别和提取。其缺省格式定义如图 3-8。

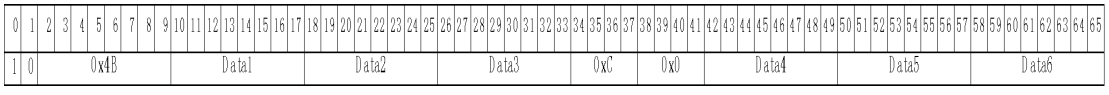
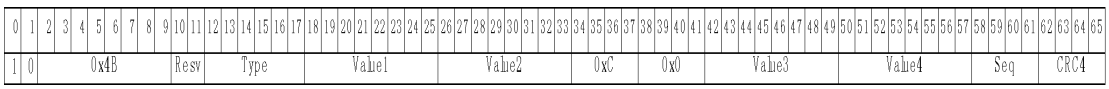


图 3-8

通道层 OAM 消息块使用 0x4B 控制码块内的第 5 个字节的高 4bits 代码区分，缺省采用 0xC，其值支持可配置。

通道层 OAM 使用控制块内的的 6 个 Data 承载 OAM 消息，使用方式如下：



具体字段格式定义如下：

- 0x4B：8bit，码块类型，表示该码块为 0 码类型；
- Resv：2bit，预留字段，缺省采用 2b00；
- Type：6bit,标识不同操作维护管理的不同功能类型；
- Value：32bit，特定类型的 OAM 消息的内容；
- C 码：4bit，缺省为 0xC，支持设置，表示 FlexE 通道层 OAM；
- Seq：4bit，标识同一 OAM 功能中采用的多个码块的序号；
- CRC4：4bit，对通道层 OAM 码块（除 CRC4 之外）的 4bit CRC 校验（同步头不参与校验）；所有操作维护管理（OAM）Block 只有在 CRC 校验正确时有效。算法多项式：X^4+x+1；初始值 0。所有操作维护管理（OAM）BLOCK 只有在 CRC 校验正确时才有效。

所有 OAM 的发送顺序与标准保持一致（注：OIF FlexE 2.0 草案 Figure-17 明确 66 比特块 bit0~65 排列，bit0 先发送，即低位在前，高位在后）。

通道层的 OAM 信息分为告警相关的 OAM、性能相关的 OAM 和其他 OAM 三大类，支持 CC、CV、REI、RDI、DM、BIP、APS、CSF、CS 等功能，详细 OAM 类型内容定义参见中移《SPN 通道层 OAM 技术方案》。

3.3.3 隧道保护

在 FlexE 通道层提供保护功能，提高客户业务在 FlexE tunnel 中传输的可靠性。当客户业务在一条 tunnel 中出现故障时，快速将客户业务倒换到另外一条 tunnel 中进行传输。保护方式分为 1+1 保护和 1:1 保护。在 1+1 保护中，可以同时在两条 tunnel 中传输，在目的点同时检测两条 tunnel 的业务服务质量状况，从服务质量高的 tunnel 中接收客户业务。

图 3-10

在 1+1 保护模式中，在发送端客户业务同时发送到两条传输通道中，在接收端检测两条传输通道的服务质量状况，从服务质量高的通道中接收客户业务。在 1+1 模式下，发送端始终处于并发状态，不需要检测管道服务质量。在接收端根据不同管道的服务质量决定选择那条管道接收，发送端和接收端独立工作，不需要传递决策信息。在 1+1 模式下，一条客户同时在两条管道中同时传输，网络承载客户业务的带宽利用率只有 50%。

FlexE 通道层保护除了 1+1 保护模式外，也可以工作在 1:1 保护模式，如图 3-11。在 1:1 模式下，有两条承载通道 tunnel：主通道 tunnel 和备通道 tunnel。在正常工作时，客户业务在主通道 tunnel 传输，备通道 tunnel 可以传输低优先级客户业务。当主通道 tunnel 出现故障时，发送到和接收端协商并决策，同时将客户从主通道 tunnel 切换到备通道 tunnel 中传输。

图 3-11

在 1:1 保护模式下，发送端和接收端需要启动 APS 协议，相互之间传递告警信息和保护决策结果，同时执行保护决策结果，保证发送端和接收端切换操作保持一致。在 1:1 保护模式下，当主通道 tunnel 工作正常时，主客户业务在主通道 tunnel 中进行传递，这时备通道 tunnel 可以传递低优先级客户业务。当主通道 tunnel 工作异常时，主客户业务在备通道 tunnel 中进行传递，占用低优先级客户业务的传输通道，这时低优先级客户业务中断。1:1 保护模式在正常工作状态下备通道 tunnel 可以传递其他低优先级客户业务，网络承载客户业务的带宽利用率可以达到 100%。

4 典型应用示例

4.1 FlexE client 业务映射示例

标准以太网业务在经过 PCS 的 64/66bit 转码(100GE)以后，其业务特征如图 4-1 所示。

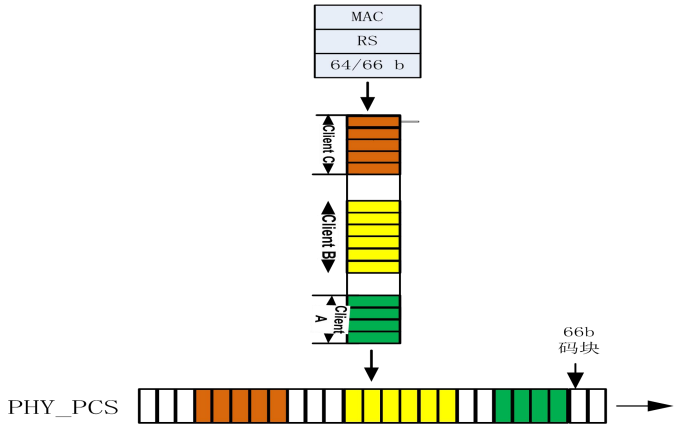


图 4-1

标准以太网接口，MAC 速率和 PHY 速率是强耦合关系，两者的带宽是严格匹配的。在 MAC 层，业务报文以字节流方式存在，业务具有突发特性；在 PHY 的 PCS 层，其承载特征是一个等效 100Gbps 速率的 66b 不间断码块流；如图 4-1，不同的业务映射到 PCS 物理层时还是以业务报文为单位的连续码块流承载。

引入 FlexE shim 功能层，实现了以太网 MAC 与 PHY 速率的解耦，其核心是在 PHY PCS 层不间断的 TDM 码块流中按固定规则插入开销码块，使得原来的无序的码块流变的可以按规则编码，把单一速率 PHY 划分成 20 个等速率载体，参见 2-7 节。

图 4-2 给出了 FlexE group 支持一个 PHY 和支持两个 PHY 的业务映射示图。在 FlexE shim 功能组件中，master calendar 对应就是 FlexE group 的时隙资源池，例如图 4-2，左边图示 FlexE group 包含一个 PHY，那么对应 master calendar 包含的资源为 20 个 5G 时隙；右边图示的 FlexE group 包含两个 PHY，那么对应 master calendar 包含的资源为 40 个 5G 时隙。对于 FlexE client，其直接可配置使用资源为 master calendar 包含的所有时隙，而不再是标准以太网中单一 PHY 定义的带宽。由此 master calendar 的应用实现了客户业务需求带宽与物理 PHY 之间的速率解耦。

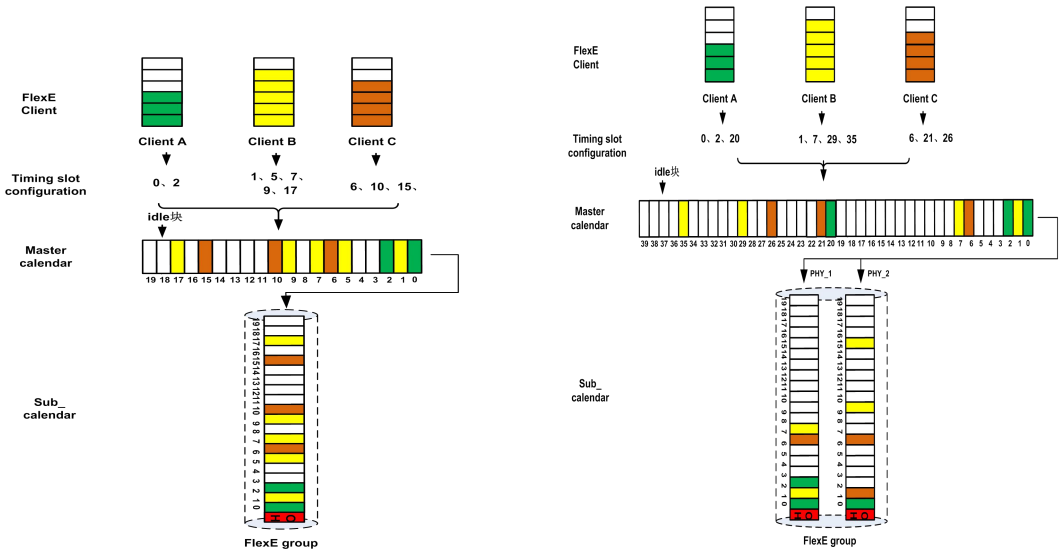


图 4-2

如图 4-2，FlexE client 可根据自身带宽需求选择配置合理的时隙数，当 client 承载的时隙指定后，该 client 对应的业务就只能在其指定承载的时隙上承载，从而实现了 client 业务承载的物理隔离。

Master calendar 的时隙通过 FlexE 段层的 PHY 来承载，其映射规则见 2.5 节的图 2-8。

同理，FlexE client 的解映射是上述业务流程的反向处理，实现机理一致。

4.2 FlexE client 交叉示例

上一章节描述了支持 FlexE 接口的 PE 节点 FlexE client 如何上下业务过程，FlexE client 交叉实现如图 4-3。

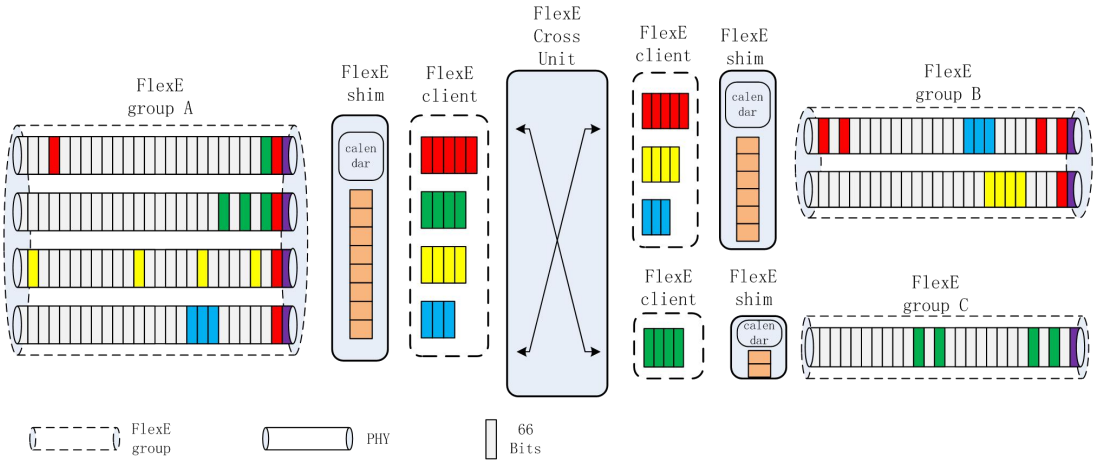


图 4-3

在图 4-3 示例的网元节点中，配置了红/绿/黄/蓝 4 个 FlexE tunnel 隧道，其中，红/黄/蓝隧道从左侧的 FlexE group A 交叉穿通到了右侧的 FlexE group B，绿色隧道交叉穿通到了右侧的 FlexE group C。如图 4-3，FlexE tunnel 端到端隧道通过 FlexE client 交叉实现，FlexE shim 通过解映射恢复出各 FlexE client 的 66bits 码块流，根据 FlexE cross unit 单元配置的连接关系，输出到对应出向的 FlexE client 单元，通过其 FlexE shim 映射到 FlexE group 发送出去，从而完成整个隧道的连通。

4.3 FlexE 技术组网应用

市场估计 5G 时代在 2020 年前开始大规模商用，基于 5G 时代移动网络的应用越来越丰富，众多应用在移动性、带宽、时延、可靠性、安全、运营计费等方面要求存在巨大差异。例如：4K/8K 移动视频业务(eMBB)要求超高带宽、超高速移动性，自动驾驶和远程控制应用（uRLLC）要求超低时延(<1ms)、高可靠性，M2M/IoT 应用(mMTC)要求超高的连接数量。每种业务类型对于网络各项能力要求非常不平均，对 5G 承载网络架构的灵活性和业务适应性提出了新的需求。如果为每种服务建立一个专用网络，成本非常高。网络切片技术可以让运营商在一个硬件基础设施中切分出多个虚拟的端到端网络，每个网络切片在设备、接入网、承载网以及核心网方面实现逻辑隔离，适配各种类型服务并满足用户的不同需求。对每一个网络切片而言，网络带宽、服务质量、安全性等专属资源都可以得到充分保证。由于切片之间相互隔离，一个切片的错误或故障不会影响到其他切片的通信。采用 FlexE 技术的网络具有弹性带宽、灵活分配的硬管道，可以实现业务的物理隔离和可靠的服务质量，天然地实现了网络切片功能。FlexE 技术的物理管道捆绑、子速率、通道化的应用模式可以承载各类速率的客户业务，提高了网络承载带宽的利用率，在 PCS 层面进行业务转发处理满足了超低延时承载的需求，降低了网络设备的成本，逐步完善的 OAM 功能满足网络维护管理需要，这些优势特点很好地满足了 5G 承载网络的技术需求。

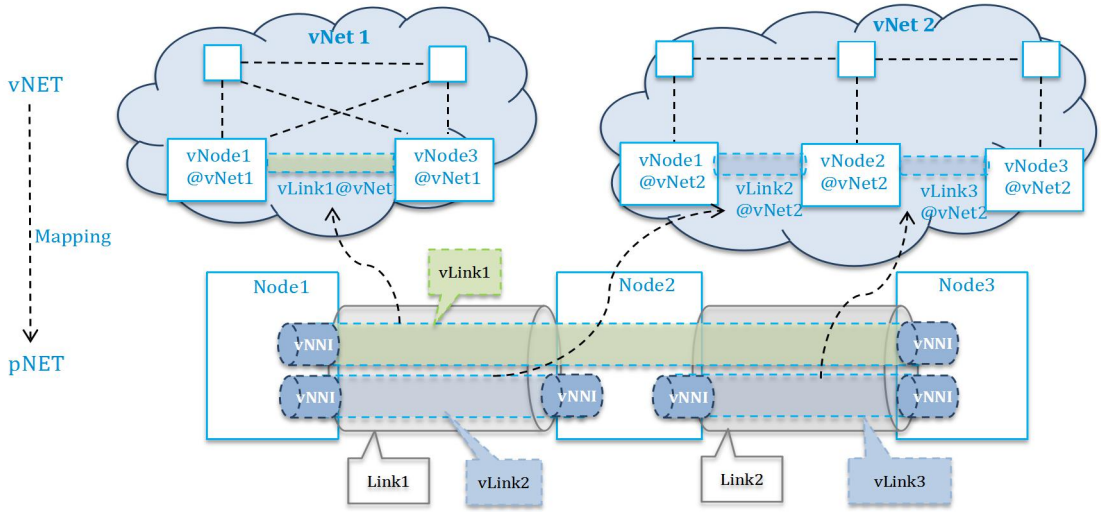


图 4-4

网络切片的实现机理如图 4-4。物理层网络节点通过 FlexE group 互联，节点网元之间配置有端到端的 FlexE tunnel 隧道，构成整个网络的互联互通管理和相关切片的逻辑连接通道，如图 4-4 中的 vLink1 & vLink2 & vLink3； vNet 虚拟切片的网络节点通过 vLink 互联，如图示的 vNet1 & vNet2；在虚拟切片 vNet1 网络中，其网络拓扑只包括与该切片业务相关的 vNode1 & vNode3，vLink1 是两虚拟网元节点的连接通道。

基于端到端物理隔离的 FlexE tunnel 隧道是后续网络切片实现的基础。

5 附 FlexE 2.0 草案介绍

说明：FlexE 2.0 目前处于草案讨论阶段，2018 年才发布正式标准。

在 2016 年 3 月发布了 FlexE 1.0 的标准内容，确定了 FlexE 协议的基础内容。在 FlexE 1.0 标准中只支持 100G 速率物理 PHY，不支持 25G、50G、200G、400G 速率 PHY。FlexE 1.0 标准中确定了管理信息传递通道，但没有确定同步时钟信息如何传送。在 FlexE 1.0 版本发布后，各厂商、组织机构开始讨论 FlexE 2.0 的内容，目前处于草案讨论阶段，最终版本在 2018 年才确定发布。从提案草案内容和多次会议结果看，有两个方向基本确定下来：

FlexE 协议中 200G、400G 速率 PHY 的承载方式和 PTP 传递通道定义。

5.1 FlexE 协议中 200G、400G 速率 PHY 的承载方式

在 FlexE 2.0 标准草案中确定了 200G、400G 速率 PHY 的承载方式：将 2 条 FlexE 成员（每个成员是 100G 速率）通过交织方式变成一条 200G 的业务流，然后在 200G 的 PHY 中进行承载；将 4 条 FlexE 成员（每个成员是 100G 速率）通过交织方式变成一条 400G 的业务流，然后在 400G 速率的 PHY 中进行承载。在 FlexE shim 层，客户业务仍按照 1.0 标准内容进行处理，master Calendar、sub Calendar 的处理过程保持不变，master Calendar 轮询分割成 n 个 sub Calendar。当用 100G 的 PHY 承载 sub Calendar 时，一个 sub Calendar 直接放在一个 PHY 中进行承载；当用 200G 的 PHY 承载 sub Calendar 时，将 2 个 sub Calendar 业务流进行交织（以 66 比特块为单位进行间插交织）形成 200G 的业务流，由 200G 的 PHY 进行承载；当用 400G 的 PHY 承载 sub Calendar 时，将 4 个 sub Calendar 业务流进行交织（以 66 比特块为单位进行间插交织）形成 400G 的业务流，由 400G 的 PHY 进行承载。如图 5-1 & 5-2，200G&400G 速率 PHY 承载方式。

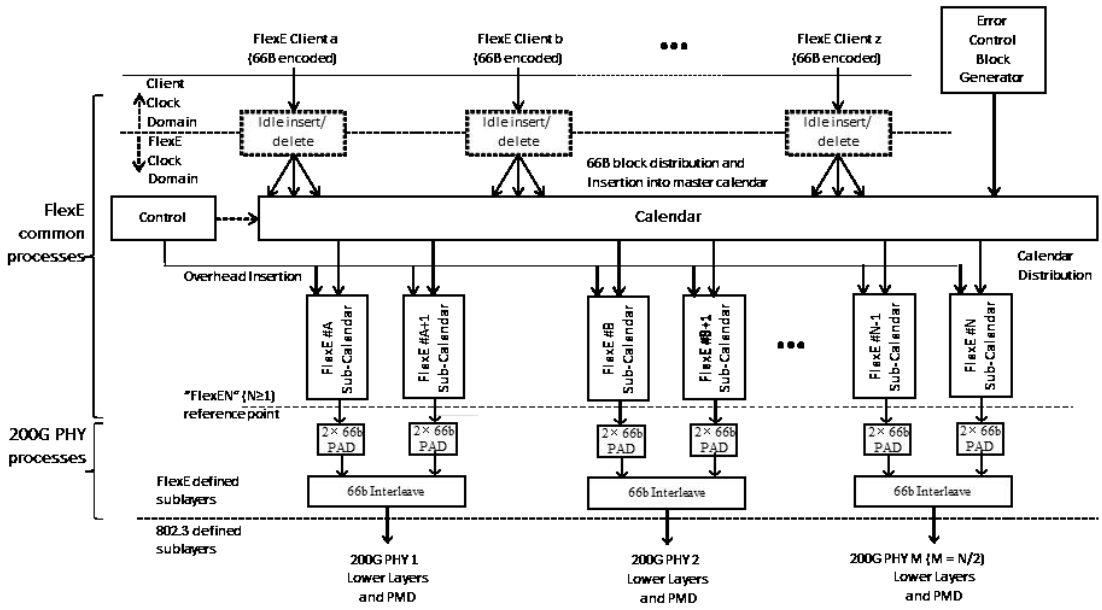


图 5-1

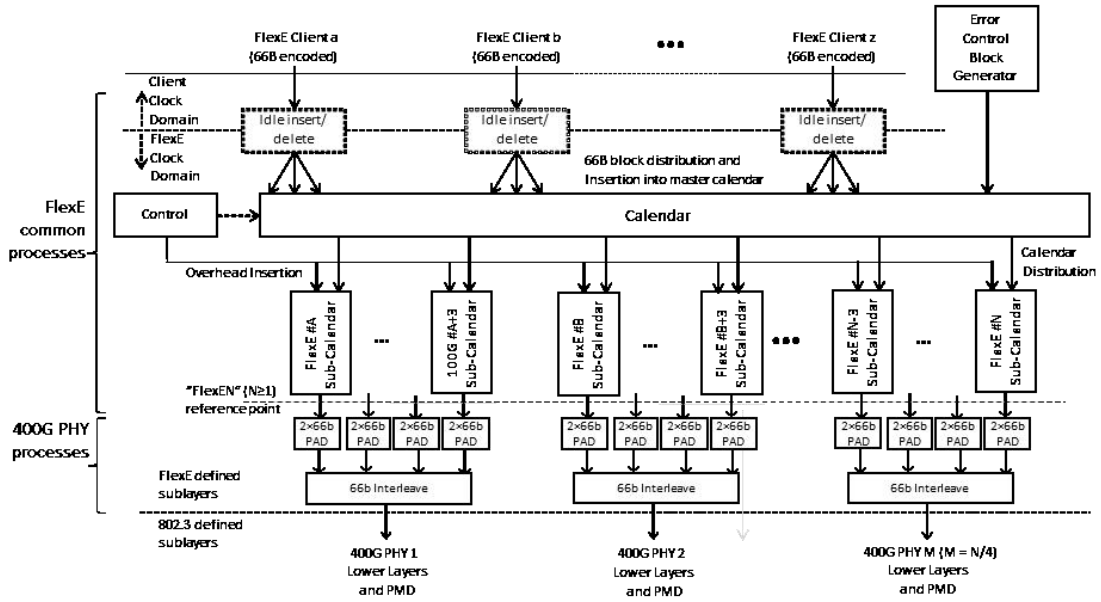


图 5-2

5.1.1 PAD 码

802.3 标准协议中，100G 速率 PHY 的 PCS 层中数据块长度是 66 比特，AM 块（多路 lane 的定位对齐标志）长度也是 66 比特，每间隔 16383 个数据块就插入一个 AM 块（AM 块出现的周期 16384 块）。但 802.3 标准协议中，200G、400G 速率 PHY 的 PCS 层，数据块的长度是 257 比特，是将 4 个 66 比特长度的数据块进行 66/257 编码后变成 257 比特长度的数据块。AM 块的长度也是 257 比特，每间隔 81916 个数据块就连续插入 4 个 AM 块，平均下来每间隔 20479 个数据块就插入一个 AM 块（AM 平均出现周期是 20480 块）。由于 200G、400G 的 PHY 中数据块的长度和 AM 块的长度都是 257 个比特，按照 257 比特长度算，AM 平均出现周期是 20480 块。如果按照比例折算，在数据块和 AM 块都为 66 比特的长度，AM 平均出现周期也是 20480 块。由于在 100G PHY 中，AM 出现周期 16384 块，相比 200G、400G 的 20480 周期，AM 在 200G、400G 中出现频率要小（相当于数据块的频率偏大）。在连续 163840 个信息块（全部折算成 66 比特块）的时间内，100G 中 PHY 中插入了 10 个 AM 块，200G、400G 中却只插入 8 个 AM 块，AM 数量变少，相当于数据块变多，导致数据块速度不匹配。为了解决数据块不匹配问题，在 FlexE 2.0 标准中对于 200G、400G 速率 PHY 承载成员业务时有特殊要求：几条 100G 的 sub Calendar 业务块流在交织前，需要先插入 2 个固定塞入 PAD 码（标称为 P1 块和 P2 块），弥补 AM 块不足的现象。

在 200G、400G 速率 PHY 中，每个 FlexE 成员先插入两个 PAD 码后，然后以 66 比特块为单位进行交织，形成 200G、400G 的业务流块，交织结果如图 5-3。

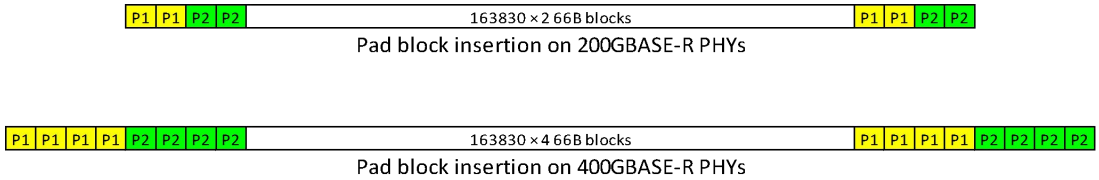


图 5-3

在每个业务流中一次插入了两个 PAD 码，标准草案中初步定义了 P1 的格式。在接收方向，根据 P1 格式可以方便地找到 P1 码。当确定了 P1 码的位置后，紧随的就是 P2 码，因此只需要定义 P1 码格式，P2 码可以是任何内容。有的厂商建议 P2 码和 P1 码格式完全一样，有的厂商建议 P2 码用来传递时钟信息、管理信息等，目前尚未确定 P2 码格式和作

用。P1 码的格式和 FlexE 帧中第一个开销块类似，通过 0X4B、0X5 来确定。但在 group number 位置，内容全部为 0XFFFFF。在 FlexE1.0 标准中规定 group number 值中 0x00000 和 0xfffff 为保留字段。在 FlexE 2.0 草案中，启用了 0xfffff 内容，规定 group number 为 0xfffff 时表示是 P1 码块。P1 码块格式如图 5-4。

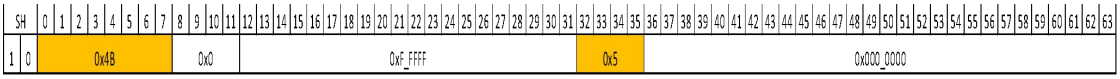


图 5-4

5.1.2 PHY number 指示

在 FlexE 1.0 标准中 PHY number 字段为每个物理 PHY 的编号，PHY number 有 8 位，最多可以标识 256 个物理成员。FlexE 1.0 标准中规定 0 和 255 两个数值为保留信息，实际应用中最多有 254 个成员。FlexE 1.0 标准规定 PHY 的速率是 100G，一个 sub Calendar 的速度也是 100G，sub Calendar 和 PHY 一一对应。在 200G 中，一个 PHY 中有 2 个 sub Calendar，每个 sub Calendar 中定义的 PHY number 无法和实际 PHY 编号对应起来。在 FlexE 2.0 中对 PHY number 字段进行重新定义，当使用 200G 速率的 PHY 承载成员时，原来字段的高 7 位表示 PHY number（其中 1-126 可使用，0 和 127 保留），最低 1 位表示 PHY 中两个实例 Instance；当使用 400G 速率的 PHY 承载成员时，原来字段的高 6 位表示 PHY number（其中 1-62 可使用，0 和 63 保留），最低 2 位表示 PHY 中四个实例 Instance。通过这种方式，每个 FlexE 成员通过 PHY number 和 Instance 两部分来表示。在实际应用中不要求一个 PHY 中所有 Instance 都有效，例如，一个 group 有 5 个成员，通过 3 个 200G 的 PHY 进行承载，最后一个 PHY 中只有一个 Instance 有效，另外一个 Instance 无效。当一个 PHY 中某一个成员无效时，通过设置该 Instance 中 group number 值为 0x00000，表示该 Instance 不属于任何一个 group 组，以此说明 Instance 无效。无效 instance 帧中第一块的结构如图 5-5。

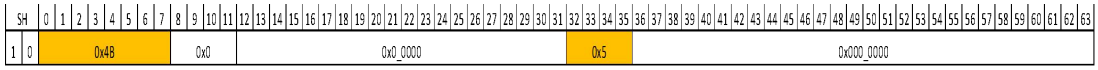


图 5-5

5.1.3 PHY 的缺陷指示方式

FlexE 2.0 草案中定义一个 200G 的 PHY 中承载两个实例 instance，400G 的 PHY 中承载四个实例 instance。在一个 PHY 中存在多个 instance，允许部分 instance 有效，部分 instance 无效。在这种情况下，草案约定：

1、 一个 PHY 中第一个 instance 必须是有效，其他 instance 可以有效，也可以无效。如果 PHY 中成员都无效，则该 PHY 就不属于 group 组。

2、 在第一个 instance 中 RPF 用来指示 PHY 的远端缺陷指示，其他 instance 中 RPF 位置值为保留信息，无任何意义。

5.1.4 PHY 速率混合应用

在 FlexE 1.0 中，要求一个 group 中所有 PHY 速度必须相同，都是 100G 速率。在 FlexE2.0 草案中定义了 200G、400G 速率的 PHY，共支持 3 种物理速率 PHY。由于 200G 速率的 PHY 中是两个 100G 业务交织成 200G 速率的业务流进行传递，400G 速率的 PHY 中是四个 100G 业务交织成 400G 速率的业务流进行传递，类似 FlexE 1.0 标准中 2 个物理通道中传递 2 个成员或 4 个物理通道中传递 4 个成员，物理 PHY 速率和承载的成员数量成比例关系，因此 FlexE 2.0 草案为不同速率 PHY 组成一个 group 提供了条件。例如，一个

500G 的客户业务，有几种不同的物理承载模式：

- 5 个 100G 速率的 PHY 组成一个 group
- 3 个 200G 速率 PHY 组成一个 group，其中一个 PHY 中只有一个 instance 有效
- 2 个 200G 速率 PHY 和 1 个 100G 速率的 PHY 组成一个 group
- 1 个 400G 速率 PHY 和 1 个 100G 速率的 PHY 组成一个 group
- 2 个 400G 速率 PHY 组成一个 group，其中一个 PHY 中只有一个 instance 有效

在 FlexE 1.0 中要求一个 group 中所有 PHY 成员全部同频。当不同速率的 PHY 组成 group 时，要求不同速率的 PHY 之间的频率成比例关系。

5.1.5 25G 速率时隙

在 FlexE 1.0 中定义 mast Calendar 是有 n * 20 个时隙组成，每个时隙是一个 66 比特块，代表 5G 的速率。随着成员数量 n 在增加，mast Calendar 中总时隙数量也在增加。在 200G、400G 速率的 PHY 中有多个 instance，当有多个 PHY 时，mast Calendar 的总时隙很多，5G 速率的单个时隙相对于 mast Calendar 总速率来讲，单个时隙的速率过小，为了方便芯片电路设计，在 FlexE 2.0 中定义了 25G 的时隙颗粒度，一个 25G 的时隙颗粒度是有连续相邻的 5 个 5G 时隙组成，并且起始时隙只能是第 1、6、11、16 等时隙，即只能是第 1、2、3、4、5 个时隙组成一个 25G，第 6、7、8、9、10 个时隙组成一个 25G，不允许类似第 3、4、5、6、7 组成一个 25G 时隙。在 Client Calendar （Client CalendarA 和 Client CalendarB）中，5 个连续时隙位置配置相同的内容，以此来表示 5 个 5G 时隙组成一个 25G 时隙。时隙速率为 25G，2 个 200G 速率 PHY 组成一个 group 的结构图 5-6。

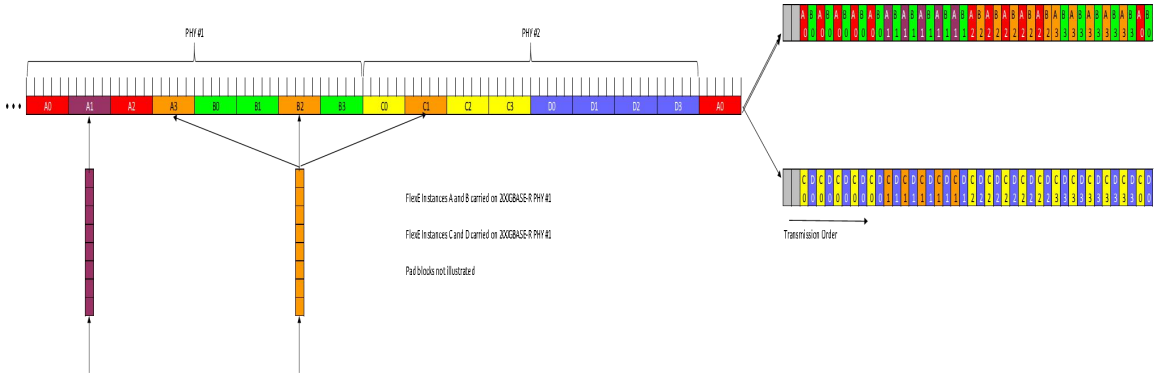


图 5-6

6 结束语

6.1 FlexE 相关国际标准组织

FlexE 标准由 OIF(**Optical Internetworking Forum：光学互连论坛**)负责制定，由于 FlexE 技术满足 5G（第五代移动技术）通讯网络的技术要求，目前许多设备厂商、移动运营商计划 ITU 标准中制定 FlexE 技术的应用标准。

6.2 FlexE 标准进展情况

FlexE 技术需求来自 Microsoft、Google 等互联网厂商，标准的主要倡导者：Juniper、Cisco、Finisar、Xilinx、NOKIA 等厂商，以及 Ixia、Deutsche Telekom、Inphi、Ciena、Kandou 等 20 多家支持厂商。FlexE 标准在 2015 年 1 月立项启动，在一年内完成草稿起草和讨论工作，迅速推出 FlexE 1.0 标准文稿，NOKIA、Inphi、Microsemi、Ciena、F7、ZTE、

Google 等公司参与 FlexE 1.0 阶段起草工作。FlexE 1.0 标准只是定义了 100G 的物理 PHY 通道，限制了 FlexE 技术的应用场景，同时 FlexE 1.0 标准中缺少同步时钟实现方案等问题，目前 NOKIA、F7、ZTE、Xilinx、Microsemi、Cisco、Juniper、Google 等多个厂商正在积极提交 FlexE 2.0 标准草稿，估计在 2018 年 2 季度推出 FlexE 2.0 标准。

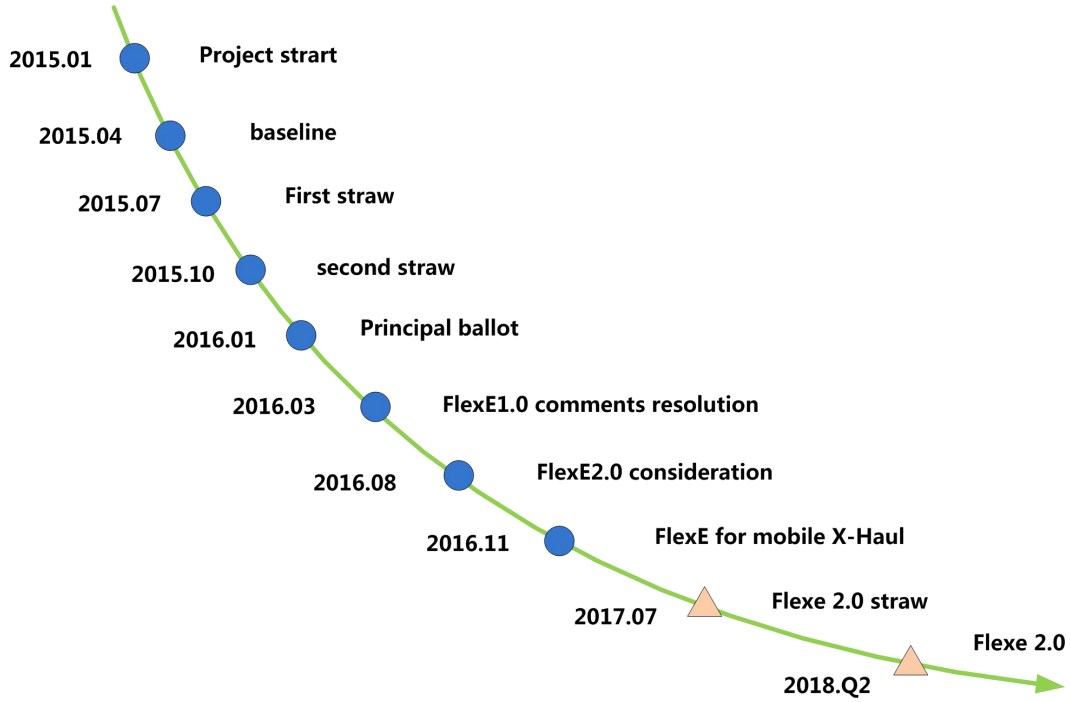


图 6-1

附录 A 参考资料

OIF-FLEXE-01.0 ： Flex Ethernet Implementation Agreement 1.0 March 2016

附录 B 缩略语

术语及缩略语表

术语/缩略语	英文	说 明
FlexE	Flex Ethernet	灵活以太网
MAC	Media Address Control	媒体访问控制
PHY	Physical layer device	物理介质
PCS	Physical coding sublayer	物理编码子层
PMA	Physical medium attachment	物理媒介接入层
PMD	Physical medium dependent	物理媒介相关层
AM	Alignment marker	定位标志
LAG	Link Aggregation Group	链路聚合组