白皮书

基于英特尔® 至强® D 处理器的扩展型皮站解决方案

作者: 刘成浴、张华哲、娄渊志

目录

1、概述

本文聚焦于如何基于英特尔® 至强® D-1747NTE 平台 (IceLake-D), 利用英特尔® 架构的各项技术, 打造中国市场的扩展型皮站产品。至强® D-1747 平台支持 5G 4x100MHz TDD 2 发 2 收端到端小区的构建以及性能测试。

本文首先介绍了至强® D-1747NTE 平台的规格和特性, 并论述了在此平台上实现扩展型皮站的能力和技术优势; 接下来描述了英特尔 FlexRAN 团队基于此平台, 构建的 4 个 5G 100MHz 2 发 2 收的端到端小区配置和测试性能, 并详尽阐述了在此平台上构建和优化上述端到端规格所采用的关键技术。

本文旨在帮助业界伙伴全面了解英特尔® 至强® D 平台的特性, 从而能更便捷和更好地基于至强® D-1747NTE 平台, 推出更具竞争力的扩展型皮站产品。

2、背景介绍

2022 年, 中国运营商的扩展型皮站集采正在进行中。英特尔针对扩展型皮站的开发需求, 不断为业界提供优异的解决方案。继上一代代号为 SkyLake-D 处理器之后, 英特尔在 2022 年又推出了全新一代英特尔® 至强® D 处理器 (代号为 IceLake-D), 其在性能、接口能力和集成度等方面都有了显著提升。为快速把新一代至强® D 处理器引入到扩展型皮站的开发中, 并为业界提供领先的解决方案, 英特尔专门定义出一款至强® D-1747NTE 平台, 并在此平台上充分验证扩展型皮站端到端系统。测试数据表明其性能可以充分匹配扩展型皮站的需求。白皮书介绍和分享了验证经验和采用的相关技术, 有助于生态伙伴共享和应用, 共同推进产业发展。

3、英特尔® 至强® D-1747NTE 平台介绍

■ 3.1 英特尔® 至强® D 平台简介

英特尔® 至强® D 处理器系列是英特尔为满足边缘计算网络、传输网络、无线网络等应用场景需求推出的 SoC (System on Chip)。其以某一代至强® 处理器为基础，集成了网络传输、加解密、存储、安全等一个或多个 IP，并根据应用场景对功耗、频率、内存通道等进行了定制，以满足边缘网络和无线网络中对处理器的多样化需求。目前至强® D 系列已经从代号为 Broadwell-DE、HewlettLake-D、Skylake-D 演进到 IceLake-D。

全新一代英特尔® 至强® D 系列 SoC (IceLake-D) 受益于新的微架构，首先带来了性能提升 (与上一代 SkyLake-D 相比)，其次具备了成熟的频率调整和功耗控制功能，且各 IP 的集成度更高，能力更贴合具体应用场景需求。

英特尔® 至强® D系列，如图 1 所示，分为 LCC (Low Core Count) 和 HCC (High Core Count) 两大类，每类又根据处理器核数量和各IP的能力包括多种具体型号。

■ 3.2 英特尔® 至强® D-1747NTE 平台的能力和特性

英特尔® 至强® D-1747NTE 是英特尔根据中国无线通信市场需求推出的一款 SoC，特别是针对中国运营商扩展型皮站的标准，做了深度定制，以便用同一平台支持最广泛的扩展型皮站需求。

英特尔® 至强® D-1747NTE 拥有10个物理核，为保证处理性能，专门提升了核心的频率和 DDR4 内存接口的速率，并集成了具备 100G 吞吐能力的以太网模块和 20Gbps 处理能力的 QAT 模块。考虑到通讯产品的工作环境，这款处理器还支持宽温，能够满足客户产品的多种部署场景要求。

英特尔® 至强® D-1747NTE 平台能力		注
核数量	10 个物理核，支持超线程 (20 个逻辑核)	
核频率	运行英特尔® AVX-512 的全核最高频率 2.9GHz，最高睿频 3.4GHz	每个核可单独调整频率
内存	3 个 DDR4 通道，最高频率 2933MT/s	
PCIe 4.0	16x	可外接用于 FEC 加速的英特尔® vRAN 专用加速器 ACC100
HSIO (High speed IO)	24 Flexible lanes	能灵活配置为 PCIe 3.0/SATA/USB 3.0
QAT	20Gb/s crypto	支持 L3 控制面和用户面消息的加解密加速
网络	共 100Gbps 的以太网传输能力，可分为 4x25Gbps 或者 8x10Gbps	集成网口
工作温度	宽温	

表 1 英特尔® 至强® D-1747NTE 平台特性

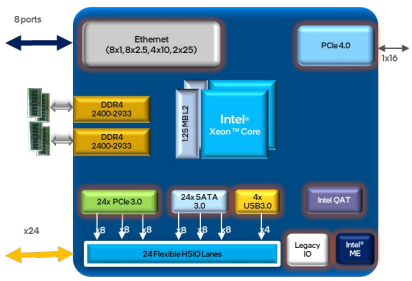
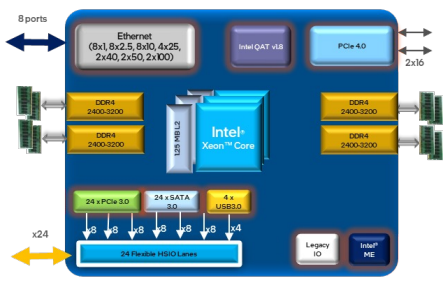
	IceLake-D LCC 规格	IceLake-D HCC 规格
核心数	2~10 Cores	4~20 Cores
DDR 带宽	支持2 or 3 通道 DDR4，最高频率 2933MT/s	支持 4 通道 DDR4，最高频率 3200MT/s
Ethernet	支持 50Gbps 或者 100Gbps	支持 50Gbps 或者 100Gbps
QAT加速器	最高 20Gbps 处理能力	最高 100Gbps 处理能力
PCIe接口	16x PCIe Gen4 + 24x PCIe Gen3	32x PCIe Gen4 + 24x PCIe Gen3
支持宽温	支持 (部分型号)	支持 (部分型号)
		

图1 英特尔® 至强® D SoC基本能力

3.3 英特尔® 至强® D 平台帮助扩展型皮站主机构建优势

3.3.1 扩展型皮站简介

随着蜂窝网络发展，以及无线通信技术的演进，异构网络逐渐成为 4G 与 5G 网络部署的主流模式。根据中国运营商发布的无线网络招投标公告，主流的 4G 与 5G 网络由宏基站、分布式小站、扩展型皮站以及飞站等几种典型站点异构而成，分别对应不同网络场景及提供不同网络服务。

扩展型皮站作为一体化皮基站的演进形态，是采用数字化技术，基于光纤或网线承载无线信号传输和分布的微功率室内覆盖方案，主要用于低容量室内场景，是室内覆盖增强方案之一。扩展型皮站由皮站主机、扩展单元、远端单元三部分组成：

- 主机单元负责基站的协议栈处理、操作维护功能，主要承载基带信号的调制和解调、无线资源管理、移动性管理、物理层处理、设备状态监控等功能；
- 扩展单元主要承载数据分路和合并、数据转发等功能；
- 远端单元主要负责射频处理及无线信号的收发。

近几年来，各运营商根据自身网络建设需求，以及通过与设备供应商的研判，逐渐形成了各自 5G NR 扩展型皮站的企业标准。这些标准虽然有所差异，但在基本软硬件功能定义和关键指标上具备共性。各设备供应商在产品规划和设计中，可考虑同平台覆盖更多的需求场景，以获得成本和供应上的优势。

3.3.2 扩展型皮站的 5G 关键功能和性能指标

各运营商扩展型皮站的部分基本软硬件功能汇总如下：

- 5G NR 扩展型皮站主机单元应支持控制面和数据面全部协议栈功能，包括层 1（L1）、层 2（L2）和层 3（L3）；
- 系统带宽方面，因运营商频段和组网需求不同，需支持子载波间隔为 30KHz 的 20MHz 以上 3GPP 定义的所有带宽；
- 帧结构方面，因运营商频段和组网模式不同，需灵活支持 5ms 单周期和 2.5ms 双周期；
- 多天线方面，需支持上行 2 天线的分集接收，每小区最大 2 流；下行支持 2 天线空分复用，每小区最大 2 流；
- 调制方式方面，下行需支持 256QAM，上行需支持 64QAM；
- 波形方面，下行支持 CP-OFDM，上行支持 CP-OFDM 以及 DFT-S-OFDM。

各设备商的方案在满足上述基本软硬件功能的同时，还需满足一些关键系统性能指标：

- 系统容量方面，至少支持 4 个 2 通道，100MHz 带宽小区的处理能力；
- 峰值速率方面，100MHz 带宽、2T2R、2.5ms 双周期帧结构，特殊时隙配比 10:2:2、下行 256QAM、上行 64QAM 配置下，应满足单载扇支持 750Mbps 下行峰值速率、281Mbps 上行峰值速率。100MHz 带宽、2T2R、74% 下行占比下，小区下行峰值吞吐量不低于 860Mbps，小区上行峰值吞吐量不低于 200Mbps；
- 传输接口方面，主机单元与核心网之间至少具备 1 个 10G 或以上速率的光接口，支持 IPRAN 的回传方式，主机单元应支持至少 4 个 10G 或 25G 的光接口，用于与扩展单元设备间的连接；
- 功耗方面，主机单元配置 4 小区 2T2R 满载工作时，最大功耗不超过 200W；
- 环境方面，应能在环境温度 -5℃～+55℃ 条件下，长期稳定可靠地工作。

3.3.3 英特尔® 至强® D 平台的优势

英特尔® 至强® D-1747NTE 平台正好满足上述扩展型皮站主机单元的关键性能指标：

- 在处理能力方面，此处理器支持 20 个逻辑核，运行英特尔® AVX-512 的全核最高睿频为 2.9GHz，并支持 3 个 2933MT/s 的 DDR4 通道，可以最大化满足 RAN 业务处理中向量化运算的需求和密集的内存访问需求；
- 在业务灵活性方面，此处理器型号集成了 QAT 芯片和网卡芯片，使用者仅需额外安装一张英特尔® vRAN 加速器 ACC100 适配器为数据信道编解码加速，其他功能均通过软件实现，从而灵活便捷地完成整个方案的调整和升级；
- 在传输方面，此处理器集成了英特尔® 以太网网络适配器 E810 网卡芯片，支持 100Gbps 的网络带宽，可以根据需要定制为不同的网口配置。例如可以配置为 4x25GbE、8x10GbE，或者 2x25GbE + 4x10GbE，能够满足扩展型皮站的组网需求；

- 在功耗方面，此处理器可以满足最大功耗要求，且具备一系列节能技术，可以进一步提升产品的能耗比；
- 在使用环境方面，此处理器支持宽温，可满足长期室外部署需求。

图 2 英特尔® 至强® D-1747NTE平台支持扩展型皮站的优势

4、基于英特尔® FlexRAN™ 参考架构的 4 小区端到端的实现及性能

英特尔 FlexRAN 团队已基于至强® D -1747NTE 平台和第三方的 RRU 与核心网，使用英特尔® FlexRAN™ 参考架构和其他第三方软件，完成了端到端的集成和性能调试，验证了此平台作为扩展型皮站主机单元的能力。下述内容会介绍这套端到端系统的软硬件配置以及基本的测试结果。

4.1 端到端系统拓扑图

基于英特尔® 至强® D-1747NTE的 BBU 端到端系统拓扑图如图 3 所示，其中：

- 居中的 BUU 框图即为基于至强® D-1747NTE 的服务器平台，安装了一块英特尔® vRAN 加速器 ACC100 适配器作为 L1 FEC 的加速，以及三条 16GB DDR4 内存条，其余均采用板载/集成的硬件和接口；
- 前传采用 4 个集成的 10GbE 接口，分别连接到 4 个 RRU 上。前传接口采用 Option 7-2，前传用户面、控制面和管理面复用同一个接口；
- BBU内部DU和CU的中传采用一个10GbE接口的两个VF端口来实现；
- 后传采用一个集成的 10GbE 接口连接到另一台 5GC 服务器，NG-U 和 NG-C 复用一个接口；
- 5GC 的 UPF 与另一台 Traffic Server 相连，RRU 通过空口连接到商用手机或测试仪表，组成了端到端系统。

4.2 BBU 平台的配置

由于不同 OEM server 开放的 BIOS 配置选项不同，以下列出了在使用英特尔开发平台测试中的关键 BIOS 配置，以供参考。

配置的基本原则是，从硬件层面上释放最大的计算能力，同时开放从 OS (Operation System) 或者用户层面控制硬件的能力，以便使硬件既能最大化满足 RAN 实时处理的需求，又通过软件控制的方式按需降频和睡眠，灵活实现节能：

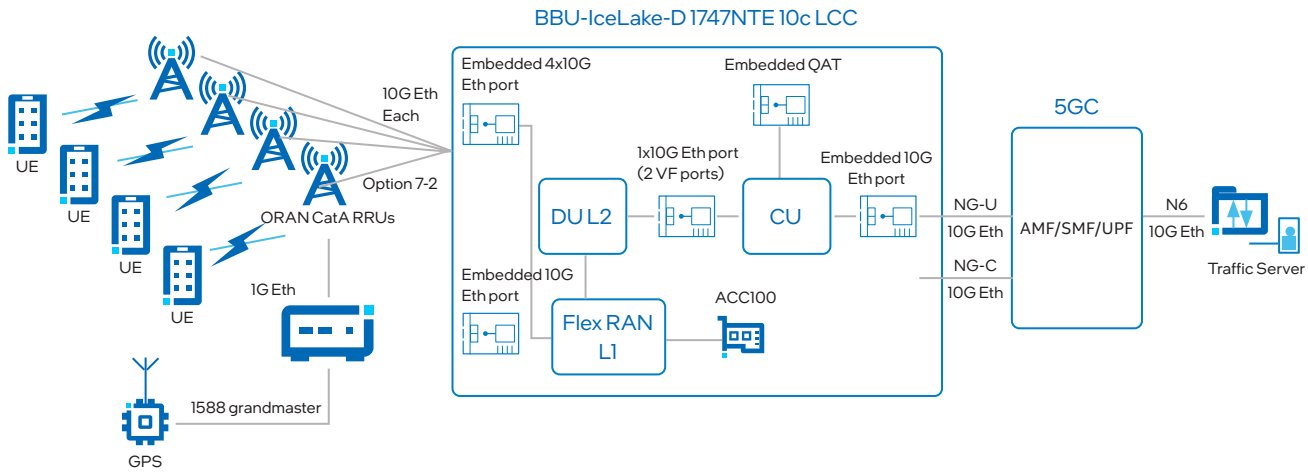


图 3 基于英特尔® 至强® D-1747NTE的 BBU 端到端系统拓扑图

[Processor Configuration] -> [Hyper-Threading All] -> Enable

- 使能超线程

[LLC Prefetch] -> Enable

- 让 CPU 可以做 LLC (Last Level Cache) 预取，减少内存读访问延迟

[Power and Performance Policy] -> Performance

- 让 CPU 运行在最好性能状态，通过 OS 和软件进行功耗控制

[CPU C State Control] -> [Autonomous Core C-State] -> Enable

- 让 OS 可以配置 CPU 的睡眠状态

[CPU C State Control] -> [Enhanced Halt State (C1E)] -> Enable

- 使能 CPU 增强型 C1 睡眠状态

[Package C State Control] -> [Package C State] -> C6(Retention)

- 让 CPU 可以进入 C6 状态

[CPU P State Control] -> [SpeedStep (Pstates)] -> Enable

- 让 CPU 可以通过 OS 配置多级运行状态，对应不同性能和功耗

[CPU P State Control] -> [AVX P1] -> Level 2

- 让 CPU 调整功耗和运行状态以支持繁重的 AVX512 处理

[CPU P State Control] -> [Turbo Mode] -> Enable

- 使能睿频

[CPU P State Control] -> [Energy Efficient Turbo] -> Disable

- 让睿频总是在功耗限定范围内以性能优先

[CPU P State Control] -> [GPSS timer] -> 0us

- 设置最小运行状态调整间隔，0us 表示可支持的最小间隔，一般为 50us

[Energy Perf BIAS] -> [Power Performance Tuning] -> BIOS controls EPB

- 设置为 BIOS 控制的 EPB 模式

[IIO Configuration] -> [Intel VT for Directed I/O (VT-d)] -> Enable

- Enable 英特尔® 虚拟化技术 (VT-d)

Grub 的配置需遵守如下几个基本原则：

- 把需要运行实时业务的核进行隔离，以避免各用户面线程互相干扰；
- 控制一些内核线程，使其不要运行到被隔离的核上，即匹配 isolcpus、irqaffinity、nohz_full 和 rcu_nocbs；
- 禁用 Intel Pstate（参照5.3.1节）驱动，避免自动调频，使每个核运行在被配置的稳定频率；
- 不使用 idle=poll，使 CPU 可以进入睡眠状态。不使用 processor.max_cstate 等参数，让 CPU 可以进入硬件允许的深度睡眠；
- 配置 iommu；
- 配置 hugepage 数量和大小。

具体 Grub 配置如下：

irqaffinity=0 default_hugepagesz=1G hugepagesz=1G hugepages=21 intel_iommu=on iommu=pt skew_tick=1 isolcpus=1-9,11-19 intel_pstate=disable nosoftlockup nohz=on nohz_full=1-9,11-19 rcu_nocbs=1-9,11-19

4.3 无线测试用例及性能

英特尔基于中国运营商部署皮站的典型 4 小区 2T2R 场景，做了一系列的测试用例。

4.3.1 4 小区 2T2R OTA 测试

我们选择 4 小区 2T2R 单用户 OTA (Over The Air) 打流的测试用例，每小区带宽 100MHz，子载波间隔 30KHz，下行 2 流，上行 2 流，TDD 帧格式为 DDDSUDDSUU。

选用此配置测试用例说明如下：

1. 小区数、流数和带宽均满足运营商企标；
2. 选用上行帧最多的 TDD 格式，以展示 L1 的处理能力；
3. 受当前测试条件限制，采用商用终端无法实现企标定义的多用户测试，所以只选择单用户打流。在此测试中，对多用户处理可能需要的资源进行了预留；
4. 受空口环境和第三方 RRU 调试原因影响，上下行速率并未到峰值，所以选择了高速率用例，但对达到峰值所需要的资源进行了预留。

在完成上述测试时，BBU 服务器的资源分配如表 2 所示：

逻辑核标号	0	1	2	3	4	5	6	7	8	9
线程分配										
逻辑核标号	10	11	12	13	14	15	16	17	18	19
线程分配										
物理核频率（GHz）	0.8	3.1	2.0	2.0	2.6	2.6	3.1	3.1	0.8	2.0

Legends:

系统核
L1 处理核（包括 timing 核和前传处理核）
L2 处理核
L3 处理核（包括后传处理核）

表 2 OTA 4 小区用例 CPU 分配、频率配置表

因至强® D-1747NTE 有 10 个物理核，打开超线程共 20 个逻辑核，所以标号是从 0 到 19。其中 0 和 10 同属一个物理核，1 和 11 同属一个物理核，以此类推。我们的配置是开启了超线程，并对同属一个物理核的两个逻辑核上运行的线程进行了配对。

IceLake 平台允许对每个物理核单独配置运行频率，两个配对的逻辑核按照物理核配置是运行在同一个频率上。在测试中，我们根据处理时延要求和节能要求对每个核频率进行了调整。

关于超线程、逻辑核线程配对、调频率等技术的具体介绍请参照第 5 章。

在实时功耗方面，整个芯片组和加速器功耗之和远小于运营商企标要求的 200W。具体功耗优化方法请参照第 5 章。

4.3.2 4 小区 2T2R Conducted Mode 测试

本测试用例使用商用的 UE simulator 来模拟每小区4用户，共4小区 2T2R 的场景，每小区带宽 100MHz、子载波间隔 30KHz、下行 2 流、上行 2 流，TDD 帧格式为 DDDDDDDSUU，前传接口和 Radio 的连接仍然是 option 7-2 接口。相比于 4.3.1 章中的测试用例，本章的用例上行速率更高，达到运营商要求的峰值速率测试用例需求。

5、英特尔® 至强® D-1747NTE 平台调优关键技术

为达到运营商企标要求的关键性能指标，英特尔 FlexRAN 团队在至强® D-1747NTE 平台上进行了系统调优，以实现功能、性能和功耗的均衡。本章节会例举其中一些关键技术和调优方法。这些技术和方法需互相配合和调谐，以达到最佳的系统性能。

5.1 英特尔® 至强® D-1747NTE 平台资源使用调优

如第 3 章所述，至强® D-1747NTE 平台集成了 QAT 和网卡芯片，可以通过灵活的配置满足扩展型皮站的不同组网需求。

在传输方面，此处理器内置了网卡芯片：

- 支持 100Gbps 的网络带宽，可以根据需要定制为不同的网口配置。例如可以配置为 4x25GbE、8x10GbE，或者 2x25GbE + 4x10GbE 等不同的配置模式；
- 前传采用 4 个集成的 10GbE 接口，分别连接到 4 个 RRU 上。前传接口采用 Option 7-2，前传用户面、控制面和管理面复用同一个接口；
- 至强® D 内置网卡支持更高精度的定时，以及 ORAN 标准的 C1/C2/C3 模式，可以作为 PTP master 给 Radio 提供授时。前传接口 Option 7-2 支持同步面和用户面控制面复用同一个接口；
- BBU 内部 DU 和 CU 的中传使用 DPDK，以提供更快速的中传用户面包处理速度。CU 和 DU 之间的 F1-U 接口可以分别采用两个 10GbE 物理接口进行回环，也可以使用同一个 10Gb E接口的两个 VF端口来实现；
- 后传采用一个集成的 10GbE 接口连接到另一台 5GC 服务器，NG-U 和 NG-C 复用一个接口。在带宽满足需要的情况下，后传的 NG-C/U 和中传的 F1-U 甚至可以复用一个物理 10GbE 接口的 3 个 VF 端口。

5.2 使用超线程提高物理核使用率

5.2.1 RAN 应用的线程分类

线程是现代计算机上最小的能被操作系统操作和调度的一系列指令流单元。一般来说我们运行的一个程序会由多个线程组成，而各线程负责相对独立的功能，通过一些控制功能实现线程间同步。

英特尔® 至强® 系列处理器采用多核架构，并且每个物理核从逻辑和算术运算的功能上看都是等效的（适用于第四代英特尔® 至强® 可扩展处理器以及之前版本的至强® 处理器）。在软件实现层面上，我们需考虑设计和实现一个多线程的应用，这样可以有效利用多核处理器进行并行处理。同时，每个线程的功能和负荷需要仔细的分配，以便使得运行这些线程的每个核得到充分利用。

但事实上一个应用的每个线程不可能均等，这意味着它们运行所需的处理器资源有差异，对时延等敏感程度也不一样。从另一个方面来说，不同线程运行在同一个核上需要的功耗也不一样。

简单给线程分类，可以分为 4 类：

- 实时线程，且包含大量 AVX 指令，典型的有 L1 数字信号处理线程，L2 的调度器线程等；
- 实时线程，但不包含大量 AVX 指令，典型的有各类 polling 线程，网络收发包处理线程等；
- 非实时线程，但需处理无线业务， 典型的有 L3 处理线程；
- 非实时线程，跟无线业务不直接相关，典型的有各类主线程、log 线程等。

以 FlexRAN L1 应用为例，各类型线程如表 3 所示：

线程类型	FlexRAN L1 线程名
实时线程，包含大量 AVX 指令	ebbupool 系列线程
实时线程，不包含大量 AVX 指令	fh_main_poll、 fh_rx_bbdev
非实时线程，需处理无线业务	无
非实时线程，不直接处理无线业务	l1app_main、l1app_stats、l1app_wlsnrt、ebbupool_main

表 3 FlexRAN L1 线程分类

对线程合理分类有助于超线程配对、调频率、线程合并等实现优化。

5.2.2 超线程基本原理和超线程配对

超线程技术能在一个物理核的基础上提供多个独立的逻辑核，使软件可以在同一个物理核上实现任务级或者线程级的并行。一般我们使用的至强® 处理器的一个物理核可以分为两个逻辑核。

这两个逻辑核各自有一整套状态寄存器，但共享一个物理核的计算资源。英特尔® 超线程技术（英特尔® HT 技术）通过管控这两套状态寄存器，让操作系统和应用与一个物理核交互中看着有两个独立的核（可参照 ***Intel® 64 and IA-32 Architectures Software Developer's Manual. 8.7 节*** https://cdrdv2.intel.com/v1/dl/getContent/671200）。此功能只需要在 BIOS 里面将 Hyper Threading 使能即可。

因英特尔® 至强® 系列是多级流水处理器，且单个指令无法同时调用所有执行单元，所以单个线程在一个物理核上执行时无法充分使用此物理核资源。而超线程可以一定程度上提高物理核资源的利用率。同时，与顺序执行两个线程相比，利用超线程可以做到两个线程并行执行，降低整体的处理时延。

为了最大化利用一个物理核内的资源，对两个逻辑核上的线程可以按一些方法进行配对：

- 可以将使用不同计算资源的线程进行配对，如将数据运算很多的线程和逻辑运算很多的线程配对。以 FlexRAN L1 为例，我们将一个 ebbupool 线程（向量运算多）和 fh_main_poll（逻辑和标量运算多）进行配对，分别绑定到4和14核上；
- 将内存访问很多的线程与内存访问不多的线程进行配对。如我们将 L3 包处理的核（内存访问多）和 L3 时钟 / 控制的线程做了配对；
- 避免将完全相同运算的线程进行配对，因为会竞争同样的资源，甚至造成性能损失。如我们将 L2 包处理的两个逻辑核进行了分散配对；
- 避免将有大量分支运算的线程进行配对，因为会竞争相同的微码缓存，降低缓存的分支路径数量。

最大化利用一个物理核内资源只是我们配对的一个准则，在实际部署中还需考虑其他准则，如：

- **需要保障某些实时线程的执行效率。**为此，可将负荷较高的实时线程（睡眠时间少）和负荷较低的线程（睡眠时间长）如非实时线程非业务线程进行配对。由此，大部分时间此高负荷实时线程实际等于独占一个物理核；
- **需在物理核数量受限情况下提高并行度，保障关键路径的时延。**比如对 FlexRAN L1，我们会打开任务切分，把一个长任务切分为一些短任务并行执行，以降低整体时延，然后依靠 ebbupool 的机制在多线程上进行任务调度。所以 FlexRAN L1 也会将一些 ebbupool 的线程进行配对；
- **节能需求。**如果处理器使能了睡眠机制，那么只有当配对的两个逻辑核都睡眠时，这个物理核才真正进入睡眠。所以将一个 polling 线程（100% 使用）跟一个处理器使用率很低的线程配对时，实际上这个物理核是 100% 在使用，并没有节能；

- **超频需求。**与节能需求相反，如果我们把一个逻辑核超频，以期获得更高的性能，实际是整个物理核都超频，功耗也会更高。那么在这两个配对的逻辑核上都可以绑定实时线程，以获得较高的能效比。关于超频可参照 5.3 节。

这些准则的运用应在满足最关键的性能指标（比如满足 L1 处理时延）的情况下，考虑可用核数量、功耗、稳定性等其他因素的平衡，并需要根据实际业务情况进行调试。

更多关于超线程的解析可以参考 ***Usage of Hyper-Threading Technology in FlexRAN™.*** *https://cdrdv2.intel.com/v1/dl/getContent/634509*。

5.2.3 线程合并以减少核占用

相比上一代用于支持扩展型皮站的 14 核至强® D-2177NT（SkyLake-D），在保证性能的前提下，至强® D -1747NTE（IceLake-D）定义 10 个物理核，其在功耗上更有优势。从至强® D-2177NT 演进到至强® D-1747NTE 平台，为充分利用新至强的单核能力，在 FlexRAN 平台升级过程中采用了如下线程合并方法：

- 优化 polling 线程逻辑，减少 polling 线程独占的逻辑核数量。我们将 FlexRAN L1 前传 polling，FEC offloading polling 和 timing polling 进行了合并，使用一个线程按定时执行轮询，既保证了轮询效率，又节省了占用逻辑核数量；
- 优化实时线程的数量。在 FlexRAN L1，所有实时业务处理，无论是需要 AVX 指令或者不需要，均采用可扩展的实时线程池。因 IceLake-D 的频率更高，所以可以灵活减少此类线程数量，从而减少所占用的逻辑核；
- 优化非实时线程逻辑核占用数量。在不影响业务的情况下，可以尽可能把非实时线程都绑定到同样的逻辑核上；
- 优化超线程的配对，尽量利用每一个逻辑核。可根据上一节所述，对配对的逻辑核上的线程进行配对，同时根据具体情况优化线程逻辑，使得所有线程都可以在超线程上运行。因 IceLake-D 频率更高，超线程配对的要求相对更低。

5.3 按需调整物理核频率

处理器每个物理核的处理能力与频率是正相关，同时越高的频率也意味着更高的功耗。至强® D-1747NTE 核频率的动态调整依托于 IceLake 平台的 P-state 技术。此技术可以对每个核进行静态或半静态的频率配置，实现性能和功耗的平衡。

5.3.1 英特尔 P-state 技术

P-state 是一个软件可见的处理器频率和性能状态指示。OS、BIOS 和任何内核线程都可以请求更改 P-state。P-state 只有在处理器运行态（C0）有效。

P-state 实际是以线程为粒度配置的。第三代英特尔® 至强® 可扩展处理器（IceLake）允许每个核运行在独立的频率上，那么每个核会根据本核上所有线程对应的最高的 P-state 运行。

通过 msr 配置工具可直接配置每核的频率（需要在 BIOS 开启 Turbo 且运行在 performance 状态），指令为"./wrmsr -p A 0x199 0xBBCC"。 "A" 是逻辑核编号，"BB" 是两位 16 进制数，表示最大的频率（以 100MHz 为单位）；"CC" 是两位 16 进制数，表示最小的频率（以 100MHz 为单位）。例如 "./wrmsr -p 2 0x199 0x1910" 表示配置核 2 运行于最高频率 2500MHz，最低频率 1600MHz 区间。如果最大频率等于最小频率，等同于期望让此核稳定到某个频率。需注意的是，处理器的功率控制器有权限决定是否采用配置，因为核频率调整既需满足单个核在最大最小频率范围内，也需所有核功耗在TDP 范围之内。

5.3.2 核频率调整带来的潜在增益

此技术非常契合 RAN 业务处理，以针对不同核所需算力升高或降低频率：

- RAN 的业务无论从纵向（协议栈）还是横向（时间）都是非均衡的：

  a、不同业务 L1/L2/L3 处理量可以完全不同，但它们往往又处于不同的线程/逻辑核上。

  b、RAN 业务存在潮汐效应，有忙时和闲时之分，对处理器的算力要求也相应变化。

- 实时和非实时线程所需的算力不一样。带 AVX 和不带 AVX 的线程可以达到的频率也不一样；
- 不但可以通过调高频率提升性能，也可以降低频率节能；
- 推荐方法是可以事先准备数套每核频率的配置，按不同场景对核频率做半静态配置。

5.3.3 按核频率调整的用例

从“表 2 处理器分配和频率配置”中可看到，我们根据多种因素，进行了线程配对和核频率调整：

- 核 1、11、6、16、7、17 调到了 3.1GHz，因为它们都是运行关键业务处理的实时线程的核，所以调到当前平台能稳定的最高频率；
- 核 4、14、5、15 调到了 2.6GHz,因为虽然核 4和 5 运行了物理层实时线程，但核 14、15 运行了 polling 线程，并不需要很高的频率，所以整体调到了一个较高频率；
- 核 2、12、3、13、9、19 调到了 2.0GHz，这些核可能是 polling 线程，也可能是负荷低的实时线程，或者是非实时线程，只需将频率调到中等即可满足运算需求，也有助于将整体功耗控制在 TDP 以内；
- 核 0、10、8、18 调到了 800MHz，这些核都是非实时线程，只需最低频率即可满足需求，运行功耗也最低。

在这种配置下，既可以满足性能需求，也可以将处理器功耗控制在合理水平。

此配置只是一种方式，建议根据业务和软件实现不同按需配置。

5.4 使用睡眠模式以节能

处理器的节能可以分为两类：运行时和非运行时（睡眠时）。运行时的功耗我们用 P-state 表征，睡眠时的功耗我们用 C-state 表征。同 P-state 一样，C-state 也是以物理核为最小单元生效的。要最大限度的节能，基本准则就是让处理器核尽量多时间处于睡眠状态，同时尽量处于最深的睡眠状态。

5.4.1 英特尔 C-state 技术

IceLake 有四种 C-state：C0（运行状态）、C1、C1E 和C6，它们的互相转化依靠 MWAIT 指令，越深的睡眠功耗越低。如果某一段时间没有任何线程调度，那么 OS 就会调用 MWAIT 指令指示此核进入某个 C-state。空闲时间越长，可进入的 C-state 就越深，当然从更深的 C-state 恢复到运行态时间也越长。一般来说从 C1 恢复到 C0 只需要 1us 以内。

5.4.2 让核睡眠的一些方法

- 首先需要 BIOS 配置“Autonomous Core C-State”，才能允许 OS 配置 C-state；
- 从 OS 的 grub 配置里去掉“idle=poll”，否则所有核不会有空闲时间，不会进入睡眠；
- 在 grub 里匹配“nohz_full”、“isolcpus”、“rcu_nocbs”和“irqaffinity”里配置的核。目的是让“nohz_full”生效（需将内核设置 CONFIG_NO_HZ_FULL=y），这样不会有周期性 tick 发生在各个核上，阻碍 OS 调取其进入睡眠；
- 在应用线程里，用“usleep”命令按需释放核，让 OS 能进行检查和调度，以使此核进入睡眠。短时间的 usleep 会让 CPU 立刻进入 C1 睡眠，更长时间的 usleep 会让 CPU 进入更深的睡眠，如 C6。进入更深睡眠的门限由 OS 决定；
- 在应用实现里，尽量减少使用内核调用函数，也尽量减少使用锁。所有内核调用也会阻止该核进入睡眠。

在英特尔® FlexRAN™ 的应用中，我们运用 eBBUPool（Enhanced BBU Pool）框架实现任务调度、分发和处理。该框架可自行根据任务量让工作的实时线程进入短睡眠，即核可进入 C1，也可通过 API 让工作线程进入长睡眠，即可进入 C6。具体参照文档 ***FlexRAN™ Reference Architecture Framework Programmer's Guide. 3.6 节*** *https://cdrdv2.intel.com/v1/dl/getContent/576898*。

6、总结和展望

本文聚焦于中国运营商扩展型皮站的关键指标，通过分析和对比英特尔® 至强® D-1747NTE 平台能力与扩展型皮站主机业务规格，验证了基于其打造的主机支持 4 个 100MHz 2T2R 5G 小区的可行性和优势。英特尔 FlexRAN 团队利用此平台搭建 4 小区端到端环境的过程，以及构建出的无线测试用例和获得的测试性能，从实践上体现了此平台的优越性能。之后，本文又对此平台的关键调优技术做了详细分析和举例，分别从超线程技术、P-state 技术和 C-state 技术描述了如何实现性能和功耗的平衡。

英特尔基于至强® D-1747NTE 平台的端到端调优还在继续，我们还在不断提高软硬件稳定性，以期达到更高的吞吐率。基于目前所示的性能数据，此平台在支持 4x100M 2T2R 后还有性能余量。我们也将继续在此平台上对 4x100M 4T4R 进行验证和调优，以期达到更高性能。

附录: 部分英文缩写及全称对照表

英文缩写	英文全称	中文全称
BBU	Building Base band Unit	基带处理单元
CU	Centralized Unit	中心单元
DU	Distributed Unit	分布单元
FEC	Forward Error Correction	前向纠错
NR	New Radio	新空口
OFDM	Orthogonal Frequency Division Multiplexing	正交频分复用
OS	Operation System	操作系统
OTA	Over-the-Air Technology	空中下载技术
RRU	Remote Radio Unit	射频拉远单元
SA	Standalone	独立组网
SoC	System on Chip	片上系统
TDP	Thermal Design Power	散热设计功耗
UE	User Equipment	用户设备
UPF	User Plane Function	用户平面功能