服务化架构 5G 核心网弹性扩容技术研究与实现

田 琪 邓子仪 罗 菊

中国电信股份有限贵州分公司云网运营部 贵州 550001

【摘 要】：针对 5G 核心网传统紧耦合架构扩容僵化、资源利用率低、不能适应多场景流量波动的缺点，本文从服务化架构（SBA）和云原生技术出发，研究弹性扩容的核心机理和实现方案。通过拆解网络功能微服务、创建动态调度模型、改善扩容触发策略，搭建起容器化弹性扩容系统，经过测试验证，该方案可以达成流量高峰秒级扩容、低谷资源回收的目的，有效地改善了网络的承载能力以及资源利用率，给 5G 核心网高效运维赋予了技术支撑。

【关键词】：5G 核心网；服务化架构；弹性扩容

DOI:10.12417/3083-5526.25.08.015

引言

5G 专用网络是指采用专有独立的无线设备 、核心网设备 ，为行业用户构建一张增强带宽 、低时延 、物理隔离的基础网络 ，实现用户数据与运营商公众网络完全隔离。当前 ，5G 专网时延优化仍面临多方面技术瓶颈 ，无线空口调度的动态适配性不足 、核心网与终端的传输路径过长 、网络资源分配与业务需求的匹配度不高等问题 ，导致时延压降效果有限且稳定性欠佳。本文聚焦 SBA 架构下弹性扩容技术难点，梳理技术框架，设计实现方案并测试验证，解决核心网动态扩容与业务保障难题。

1 5G 核心网服务化架构与弹性扩容需求分析

1.1 服务化架构（SBA）核心特性

5G 核心网抛弃了 4G 传统点到点网元耦合架构，依照 3GPP R15 和后续版本所规定的服务化架构，把核心网控制面的功能拆解成 AMF、SMF、UDM、PCF、NRF 等独立的网络功能服务（NF Service），各个服务借助标准化的服务化接口（SBI）按照 HTTP/2 协议互相交流，从而达成功能解耦、独立部署以及灵活调用的目的。相比于传统的架构，SBA 架构有三个主要的优势，分别是微服务化拆分、控制用户面分离(CUPS)和云原生适配。用户面 UPF 和控制面服务完全解耦，可以分别按需扩容，满足大带宽、低时延业务的需求，云原生适配，天然支持容器化、虚拟化部署，支持基于容器编排工具的动态资源调度，给弹性扩容提供架构底层支持[1]。

1.2 传统扩容模式的局限性

4G 核心网和早期 5G 核心网大多采取静态扩容方式，也就是依照业务峰值预估来安排硬件和软件资源，日常运作时资源被固定分配，不能随意变动。该模式存在很多不足之处，其一，资源利用率很低，峰值配置造成闲时大量的算力、带宽资源被浪费，运营商运维成本居高不下；其二，扩容响应迟缓，静态扩容需要人工干预、硬件部署和配置调试，耗时数小时甚至数天，不能应对节假日、大型活动、工业现场突发流量；其三，扩容粒度过粗，传统的网元整体扩容不能对高负载的单一服务组件进行精确的扩容，容易造成局部资源瓶颈；第四，业务连续性差，在扩容过程中很容易造成业务中断、信令风暴等问题，影响用户体验以及行业应用的可靠性。

1.3 弹性扩容核心需求与技术指标

根据 5G 多场景业务特点和 SBA 架构特点，核心网弹性扩容要达到四个主要要求，即实时性、精准性、自愈性、经济性。实时性，流量变化的时候可以做到秒级的资源调配，不会造成业务拥塞；精准性，支持微服务粒度的精细化扩容，针对局部负载的压力进行有针对性的缓解；自愈性，在出现故障的时候能够自动扩充冗余节点来保证业务的连续运行；经济性，低谷时期自动缩小容量释放资源，削减能耗和运维费用。核心技术指标有以下两个方面，扩容响应时间≤5 秒、服务实例的扩缩容粒度是单个容器实例、资源利用率>65%、业务中断率 0、适配 AMF、SMF 等核心控制面服务和 UPF 用户面功能的不同扩容需求[2]。

2 服务化架构 5G 核心网弹性扩容关键技术

2.1 基于云原生的微服务容器化封装

弹性扩容的基础就是核心网功能的容器化改造，把 SBA 架构下各个网络功能服务拆解成无状态微服务组件，去除状态依赖，用 Docker 容器来完成封装。容器化实现了服务和底层硬件的解耦，具有轻量、启动快、迁移灵活的特点，相比传统虚拟机来说，容器启动时延从分钟级降到秒级，满足了弹性扩容的实时性要求。构建统一的容器镜像仓库，对各个服务版本进行标准化管理，防止扩容过程中版本不一致造成业务异常，对于无状态的控制面服务（AMF、SMF 等），进行无状态化改造，支持多实例并行部署和负载均衡，为水平扩容打下基础，对于 UPF 用户面功能，优化容器网络性能，保证数据转发时延和吞吐量。

2.2 动态负载感知与扩容触发策略

准确的负载感知是弹性扩容的前提，本文设计出多维度的负载感知模型，实时采集核心网微服务组件的 CPU 利用率、内存占用、网络带宽、信令处理速率、并发连接数等核心指标，

使用 NRF 服务注册发现功能获取各个服务实例的运行状态和负载分布。抛弃传统的单个阈值触发方式，创建起复合阈值加上趋势预估的双层触发体系，基础层确定 CPU 大于等于 80%，并发连接数超过某个阈值这些硬性扩容标准，预测层借助历史流量数据，利用时序预测算法来预估流量高峰，提前发出预扩容指令，从而防止因突然出现的流量引发的网络拥堵状况。另外还设置了缩放延时，防止因为短时间内流量的变化而造成频繁的扩缩容，保证系统的稳定[3]。

2.3 容器编排与动态调度技术

使用 Kubernetes(K8s)作为主要的容器编排平台搭建 5G 核心网的弹性调度集群，对服务实例进行自动化部署、扩容、缩容和故障迁移。基于 K8s 的 Horizontal Pod Autoscaler（HPA）机制，定制适配 5G 核心网业务特性的扩缩容算法，支持根据自定义业务指标动态调整 Pod 副本数。根据核心网服务优先级的不同来制定资源调度优先级策略，保证 AMF、UDM 等核心控制面服务先于其他业务获得资源，uRLLC 低时延业务扩容优先于普通 eMBB 业务。结合亲和性与反亲和性调度策略，防止同类服务实例过于集中地部署在同一个节点上，防止单点故障造成整个业务系统瘫痪，达到跨节点、跨区域的分布式弹性扩容的目的。

2.4 服务化接口适配与负载均衡优化

SBA 架构下的各个微服务之间是通过服务化接口进行交互的，扩容之后新增的服务实例要迅速加入到集群当中，并且还要保证流量能够得到均衡的分配。通过部署服务网格（Istio）组件来对微服务之间的流量进行精细化的控制，新增实例会自动完成服务注册和发现，NRF 实时更新服务目录，使调用方可以很快地知道新增加的实例。使用四层加七层的负载均衡方式，四层负载均衡做流量初步分发，七层负载均衡针对 HTTP/2 协议做会话保持和流量精细化调度，防止单个实例超载。同时对服务调用链路进行优化，扩容后会自动均衡各个实例的负载，无状态服务可以无缝接入，整个过程不会影响到现有的业务运行，实现了无损弹性扩容。

3 弹性扩容系统设计与实现

3.1 系统整体架构设计

本文所设计的SBA架构5G核心网弹性扩容系统由上到下分为业务感知层、负载监控层、策略决策层、容器编排层、基础设施层五个层次。业务感知层与 5G 核心网业务平台对接获取各场景业务流量及服务调用数据，负载监控层采用 Prometheus 监控组件对各个微服务实例性能指标进行实时采集、清洗和分析，策略决策层为业务扩展与收缩触发模块，根据负载感知数据发出扩容或者缩容命令以完成扩缩容任务，容器编排层采用 K8s 集群对实例执行扩缩容、调度以及部署操作，支撑虚拟化和容器化的应用，基础设施层为服务器、存储、网络等硬件资源赋予支持，保证系统的虚拟化和容器化运转。各个层次之间用标准的接口进行交互，从而达到全流程的自动化闭环，无需人工操作。

3.2 核心模块实现

3.2.1 负载监控模块

根据 Prometheus+Grafana 创建可视化监控平台，创建出针对 5G 核心网的专用监控指标，包含控制面信令处理、用户面数据转发、系统资源占用这三个方面共 20 多个指标，使用 Exporter 组件采集容器和服务实例的数据，把数据采集频率设为 5 秒，保证负载感知的及时性。根据核心网服务分布式部署的特点，在增加跨节点数据聚合功能的基础上，将各个区域的服务实例数据进行整合，消除由于数据碎片化造成的监控盲区，改善数据传输协议，减少监控组件自身所占资源，防止监控模块对核心业务造成影响。监控平台支持阈值告警以及趋势可视化展现，运维人员可以随时了解到各个服务的负载状况，并且可以把监控数据立即发送到策略决定模块中，从而给扩缩容决策赋予准确且即时的数据支持。

3.2.2 扩容决策模块

采用 Python 开发出一套自定义的决策算法，将实时监控的数据和流量预测模型结合起来，设置三级扩容策略，一级轻量扩容，单服务指标轻微超标，增加 1 到 2 个实例；二级中度扩容，指标持续超标，增加 3 到 5 个实例；三级重度扩容，突发流量峰值，增加 5 个以上的实例，并且触发跨节点调度。算法自带流量趋势拟合功能，用历史 72 小时流量数据做短期预测，区分正常潮汐波动和突发流量，防止误判造成无效扩容。缩容策略使用梯度递减的方式，流量下降的时候分批缩减实例，防止一次性缩容造成二次扩容，设置 10 分钟的缩容冷却期，过滤短时流量波动。决策模块同 K8s API 完全对接，可以完成指令的秒级下发和结果的实时反馈，整个过程不需要人工干涉，实现了全流程的自动化执行[4]。

3.3.3 容器编排与服务部署模块

完成 5G 核心网核心微服务容器化镜像制作，优化容器网络配置，使用 Multus CNI 实现多网卡容器部署，满足核心网控制面和用户面分离的网络需求，单独划分控制面信令通道和用户面数据通道，避免两类流量互相干扰。对核心网服务镜像进行轻量化裁剪，去除无用的组件，减小镜像大小，加快实例启动和扩容部署的速度。定制 K8s 调度策略，根据核心网服务无状态的特点来实现服务实例的水平扩展，配置 HPA 自动扩缩容规则，绑定自定义业务指标和系统资源指标，打破原生 HPA 只用资源指标的局限。同时加入故障自愈功能，服务实例出现异常的时候会自动销毁新的服务实例，并重新创建起来，配合弹性扩容的功能，实现对流量和故障双重驱动的资源动态调节，从而保证核心网的服务连续性。

3.3 系统部署与业务适配

系统部署在运营商现网测试环境，使用 6 台物理服务器组成 K8s 集群，完成 AMF、SMF、UDM、UPF 等核心网功能服务的容器化部署，对接模拟基站和终端设备，模拟出三种典型的业务流量，分别是 eMBB 高清视频、uRLLC 工业控制、mMTC 物联网，模拟出现网业务流量模型及网络拓扑，保证测试结果具有现网参考价值。在部署的过程中严格按照运营商现网的安全规范，设置网络防火墙以及访问权限，防止测试环境对现网业务造成影响。根据不同的业务场景来设置不同的扩容参数，eMBB 场景重在带宽和并发连接数的指标上，扩容响应速度较快；uRLLC 场景重在低时延保证上，扩容时先保证转发性能，留出专属算力资源；mMTC 场景重在连接数承载上，优化海量终端接入时的扩容效率，实现多场景弹性扩容精准适配，满足 5G 全业务场景的需求。

4 系统测试与结果分析

4.1 测试环境与方案

为了检验系统的性能，搭建了传统的静态部署组和本文所提出的弹性扩容组两个测试环境，两组的硬件配置、网络环境完全相同。测试场景分为三类，分别是突发流量场景、潮汐流量场景、混合业务场景，分别模拟大型活动期间 eMBB 业务突发流量、日常早晚高峰流量波动、三大 5G 典型业务并发。测试指标有扩容响应时延、资源利用率、业务中断率、信令成功率、数据转发时延，每组测试做 3 次取平均值。

4.2 测试结果与对比分析

从测试结果可以看出，本文所设计的 SBA 架构弹性扩容系统比传统的静态扩容模式有较好的性能。突发流量场景下扩容响应时延为 3.8 秒，是传统模式的 2 小时以上，系统平均资源利用率由原来的 32%提高到现在的 68%，大大减少了资源闲置损耗，潮汐流量场景下早晚高峰自动扩容、低谷自动缩容，全程无人工干预，业务中断率为 0，信令处理成功率保持在 99.99%以上，混合业务场景下可以对不同的服务进行精确的扩容，uRLLC 业务转发时延稳定在 10ms 以内，满足低时延要求，mMTC 场景下支持 10 万级终端并发接入，没有出现拥塞情况。系统运行稳定，没有由于频繁扩缩容造成的服务抖动，完全满足 5G 核心网运维的要求。

4.3 问题与优化方向

测试中发现有状态服务组件的扩容适应性较差，跨区域扩容的时候数据同步会出现短时间延迟的情况。为了解决上述问题，可以对状态数据缓存和同步进行改进，使用分布式缓存技术来达到扩容实例之间数据快速同步的目的，并且可以进一步改进流量预测算法，加入 AI 深度学习模型来提高预测的准确性，从而达到更早的智能预扩容的效果，进而减小突发流量给网络带来的影响。

5 结语

本文从 5G 核心网业务发展的痛点出发，利用服务化架构和云原生技术来解决弹性扩容的技术问题，从而解决了传统架构扩容僵化、资源浪费的问题。本方案可以有效地适应 5G 各种场景的流量变化，提高网络的可靠性和资源利用率，有现网部署的价值。可以利用 AI 智能调度、6G 预演进技术来达到全域弹性协同、跨网联动的目的，促使核心网朝更智能、更敏捷、更高效方向发展。

参考文献：

[1] 黄灏.基于微服务的无线接入网用户面架构研究[D].北京邮电大学,2025.

[2] 许经广. 工业互联网平台架构设计与关键技术研究[J].无线互联科技,2025,22(08):106-109.

[3] 赵毅鹏. 微服务化的即时通信系统架构分析与设计[J].机电信息,2025,(08):40-43.

[4] 关玉莲. 无线电监测应用系统的微服务化[J].中国无线电,2024,(12):51-53+56.