国家高性能计算环境建设

赵一宁

zhaoyn@sccas.cn

中国科学院计算机网络信息中心

国际主要高性能计算环境

EGI
- 欧洲网格基础设施
- 欧洲各国家网格联合而成
- 前身为EGEE

XSEDE
- 极限科学与工程发现环境
- 美国院校与研究机构资源整合
- 前身为TeraGrid

WLCG
- 世界LHC(大型强子对撞机)计算网格
- 世界范围计算环境
- 专门用于高能物理试验计算

国际主要高性能计算环境

	XSEDE (TeraGrid)	EGI (EGEE)	WLCG	CNGrid
启动时间	启动：2001 本期：2011	启动：2001 本期：2010	启动：2002 运行：2006	启动：1998 本期：2016
组织主导方式	国家主导	机构合作	机构合作	国家主导
资源所属	美国国内	全球范围 欧洲为主	全球范围	中国国内
站点数	19	300+	170+	19
聚合计算资源	16.1PF			200PF
聚合存储资源	86.4PB			160PB
中间件	Globus	EMI（ARC、gLite、UNICORE、dCache）	EMI（ARC、gLite、UNICORE、dCache）	SCE
用户数	~5500	>10000	~9000	~10000

	XSEDE (TeraGrid)	EGI (EGEE)	WLCG	CNGrid
主要资助方	美国国家科学基金会（NSF）	欧洲委员会（European Commision）	与EGI互相结合	中国科技部
项目资助时间与额度	2001年-5300万美元 2002年-3500万美元 2003年-1000万美元	2001年-1200万欧元		2002年-1亿RMB
	2005年-1.5亿美元	2004年-4600万欧元 2006年-5260万欧元		2006年-9.4亿RMB
	2011年-1.21亿美元	2010年-3200万欧元+各国国拨资助		2011年-13亿RMB
	2016年-1.1亿美元	2016年-3700万欧元 2017年-8600万欧元		2016年-
主要应用领域	生物及交叉学科 神经科学 物理学 化学 材料科学 天文学科学 分子学相关学科	高能物理 生命科学 天文学与天体物理学 计算化学与材料科学 地球科学 核聚变	高能物理	量子化学 分子模拟 高能物理 生物科学 流体力学 材料科学 大气与海洋学 天文学

数据来源： *https://www.xsede.org* *https://ec.europa.eu* *https://en.wikipedia.org*

国家高性能计算环境

始终坚持高效能计算机、高性能计算服务环境和高性能计算应用三位一体、均衡发展的战略。以高效能计算机提供基础计算资源，以服务环境实现资源共享，降低应用门槛，以应用的发展促进机器和环境的技术进步。

研制
超级计算机

- 天河一号、二号
- 神威·太湖之光
- E级超级计算机
- 自有技术，国际领先

发展
高性能计算应用

- 超大规模并行应用
- 材料、生物医药、复杂工程、环境等应用领域需求
- 奠定我国的计算软件产业基础

构建
高性能计算环境

- 解决运维和用好超级计算机的问题
- 建设新型国家基础设施
- 催生我国计算服务业

发展历程

国家科技重点研发计划

863重大专项
高效能计算机及应用服务环境

863重大专项
高效能计算机及网格服务环境

863重大专项
高性能计算机及其核心软件

中国科学院信息化建设专项-超级计算环境建设

十五 → 十一五 → 十二五 → 十三五

- 2005.12 中国国家网格运行管理中心挂牌
- 10个结点

- 国家超级计算中心相继成立
- 11个结点

- 2013.9 超级计算创新联盟成立
- 14个结点

目标：

- 资源进一步扩展
- 高速主干网络
- 海量用户数据存储
- 环境实时展示
- 环境服务化运行

国家高性能计算环境 CNGrid

双运行中心（北京/合肥）

19个结点（200PF+162PB）

互联带宽1000Mb
（北京/合肥/无锡/广州/上海）

基于应用的全局调度与预测

基于微服务结构的计算门户

院超级计算环境

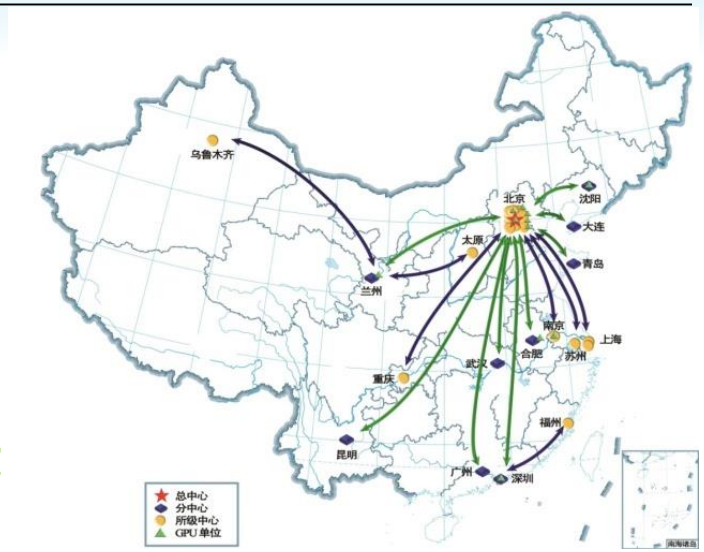
合肥分中心验收通过

兰州分中心验收通过

青岛分中心验收通过

昆明分中心验收通过

大连分中心验收通过



2010.9.21 2010.10.09 2010.10.13 2010.10.27 2010.11.03 2010.11.29 2011.01.05 2011.03.11 2009.6-2014.12 2015.7.2 2015.7.16

19家所级中心
完成协议签署

深圳分中心验收通过

沈阳分中心验收通过

武汉分中心验收通过

合肥分中心（中科大）验收通过

广州分中心验收通过

环境核心系统软件SCE

从零开始　2005年起探索自研SCE系统软件

第一阶段

2009年12月开始部署在院超级计算环境
主要面向科学计算服务
提供方便的命令行/Portal等多种形式

第二阶段

2013年11月发布环境应用编程接口
平台无关、语言无关、简单易用
基于接口支持10+领域云/专业社区

第三阶段

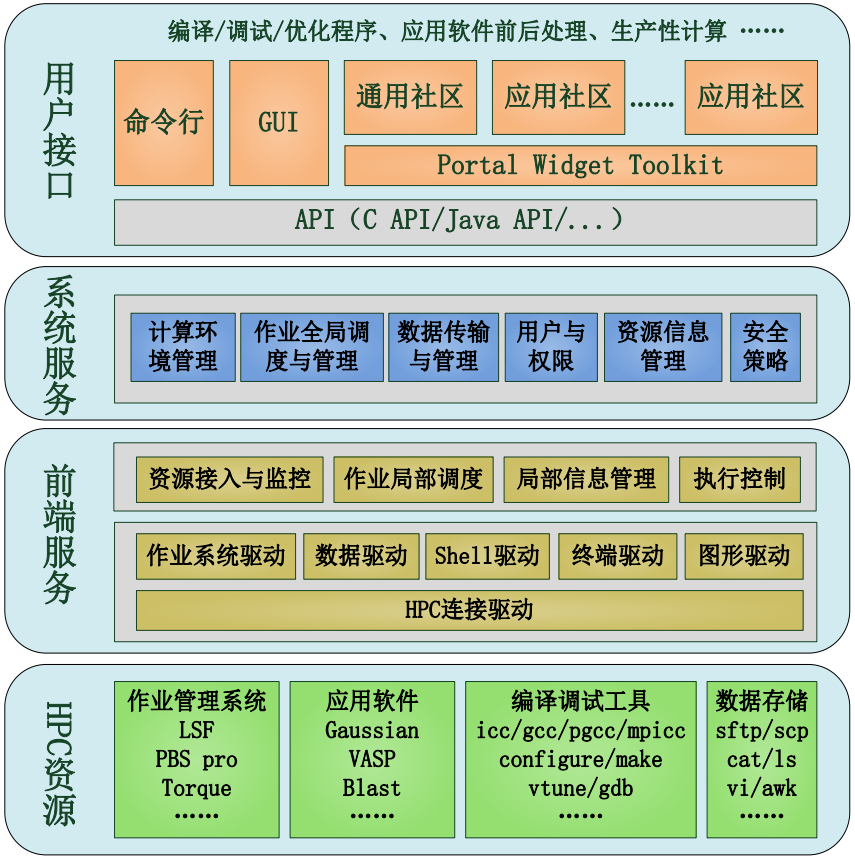
2014年7月用于构建国家高性能计算环境
“可运行、可管理、可使用”
监控和运行管理平台/软件快速交付

全面发展

系统软件结构优化
事件流平台构建
环境安全保障提升

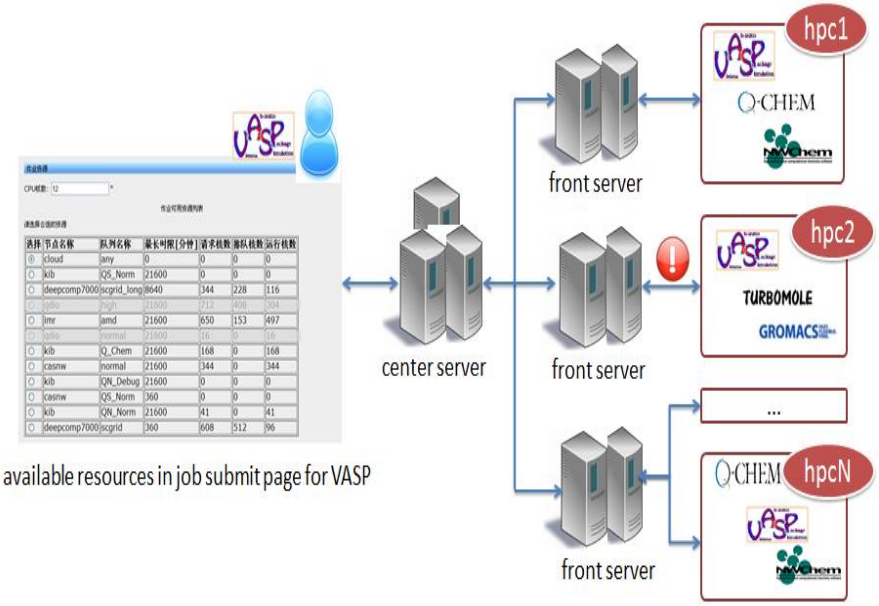
环境核心系统软件SCE

- SCE软件
  - 聚焦高性能计算服务
  - 稳定、可靠、维护成本低
  - 两种使用方式
    - Portal和命令行
  - 2009起稳定运行至今
  - 申请PCT国际专利1项



全局资源调度

- 以应用为中心的资源弹性调度
- 完善超级计算资源的优化配置与利用
- 屏蔽集群异构性（硬件/编译环境/执行环境）
- 作业在被调度的资源执行最优，兼顾排队时间



available resources in job submit page for VASP

资源收集器

- 每5min收集一次队列信息
- 定期维护应用信息
- 按需求维护用户映射信息
- 按需求维护资源权限信息

资源匹配器

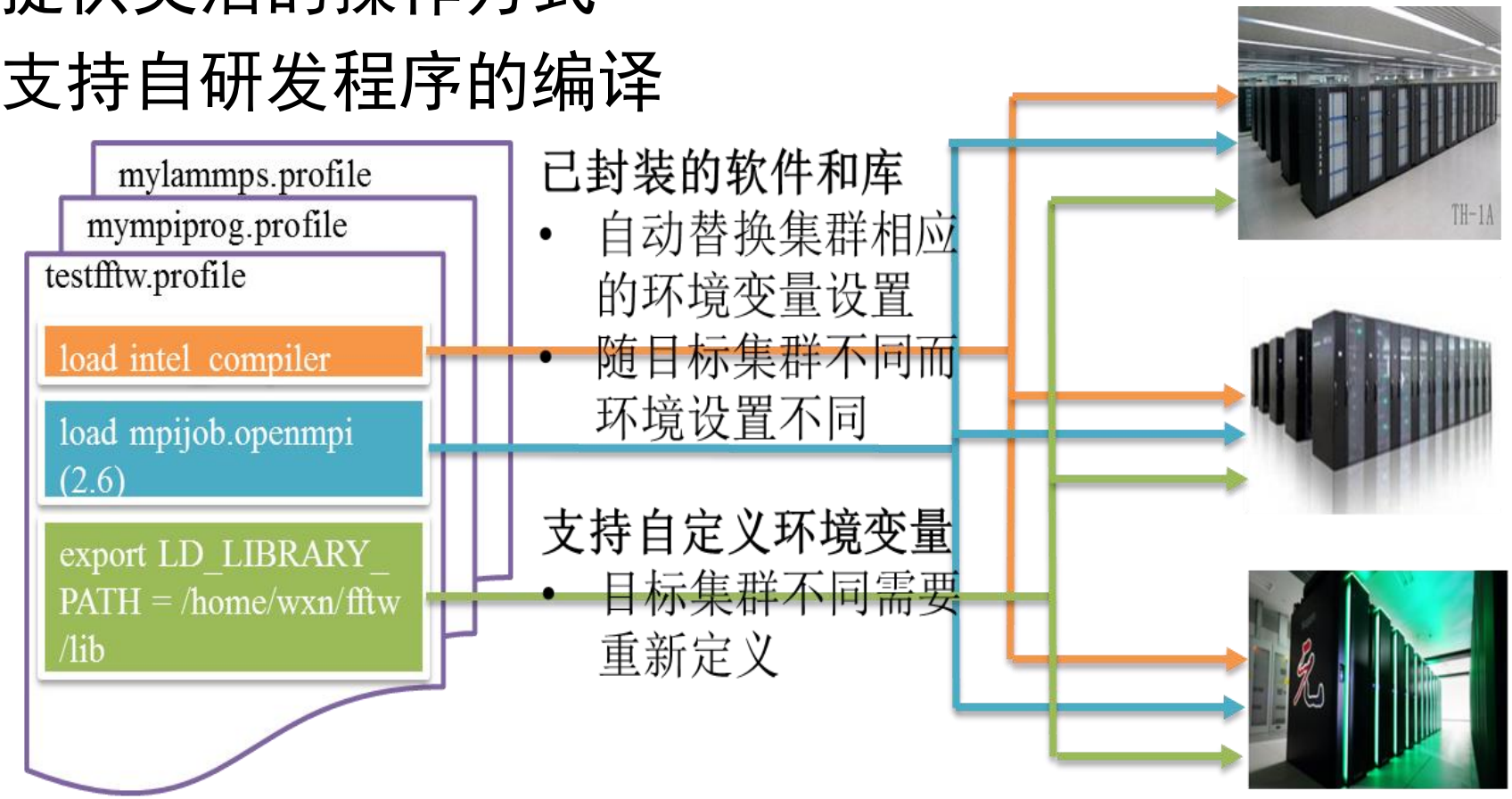
- 匹配集群和队列资源
- 匹配应用和应用版本资源
- 匹配核数限定和执行时间限定
- 匹配用户映射和资源权限

资源调度器

- min{排队核数×作业执行时间限定}
- 结合应用软件在特定集群的已有测试结果或使用经验
- 增加网格队列，避免非用户原因引起的提交出错

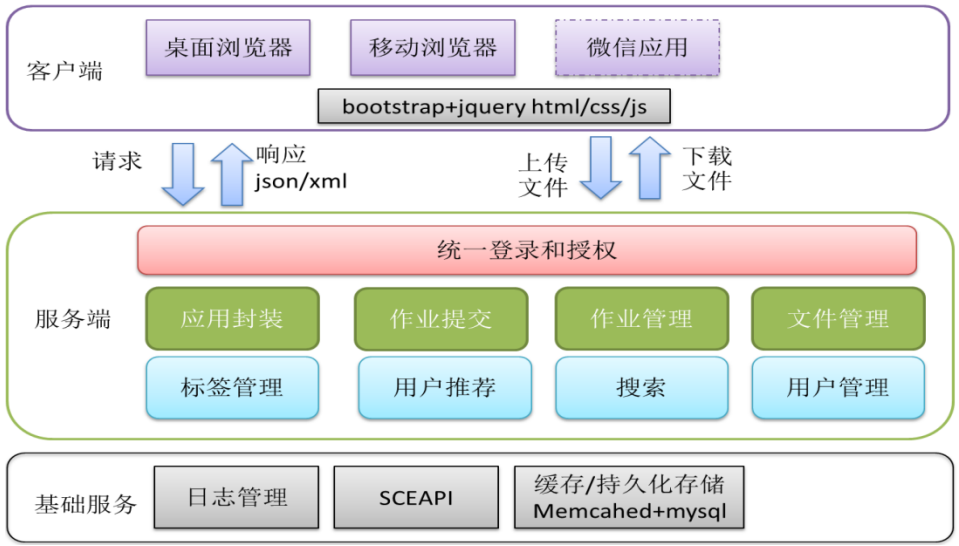
命令行使用方式

- 面向传统用户
- 提供灵活的操作方式
- 支持自研发程序的编译



计算门户Portal 2.0

- 从部署到用户体验的全新蜕变
  - 微服务理念
  - Docker部署
  - 多实例部署
  - 资源标签化管理
  - 智能搜索
  - 资源关联推荐



- 目前进展：
  - 第一版（已发布）：在原有作业提交和查看的功能基础上用户操作更加顺畅，提示更加详细
  - 第二版（已发布）：优化微服务结构，新增英文版
  - 第三版（研发）：用户信息维护模块，更多应用封装
  - 更多计划（原型实现）：完成Portal2.0形成资源搜索的整体技术方案并正在原型实现，形成相关文档

环境应用编程接口

- 环境应用编程接口
  - RSET风格
  - 跨平台和跨语言
- 设计原则
  - 所有的资源都抽象为URL的集合
  - HTTP动作
    - GET：幂等操作，查询资源
    - POST：创建或更新一个资源
    - PUT：更新一个资源
    - DELETE：删除资源
- 已支持10+个应用社区和领域云建设

环境监控与运行支持平台

用户管理

申请审批流程更为便利顺畅

技术支持

作业差错更加安全高效

集群管理

环境层面资源访问权限可控

环境展示

资源使用、服务异常尽在掌握

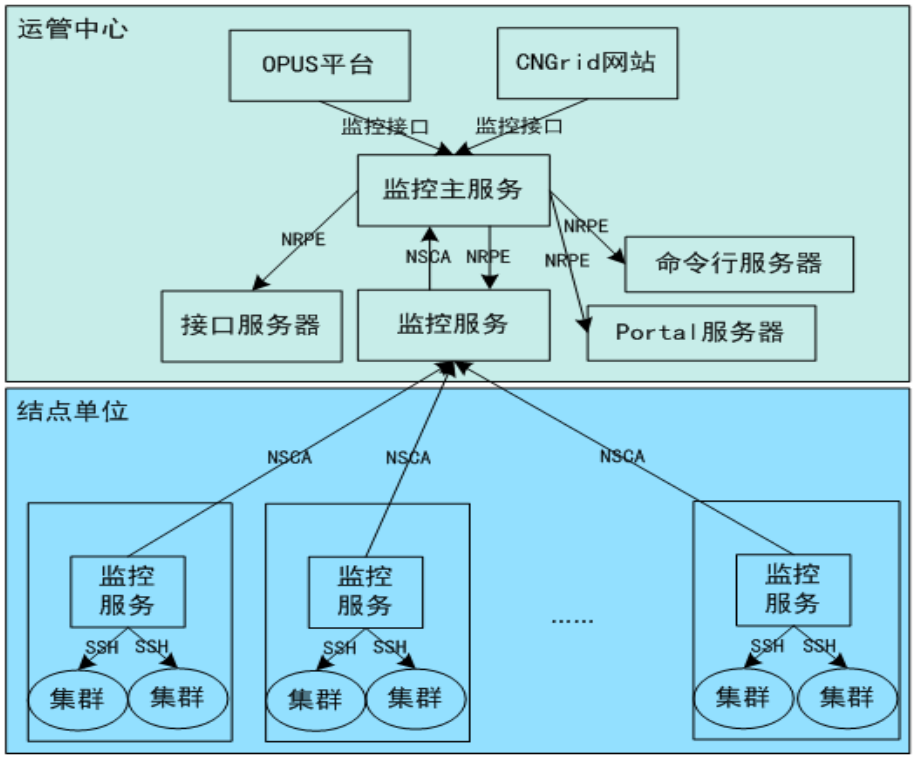
作业统计

支持各类统计数据下载查看

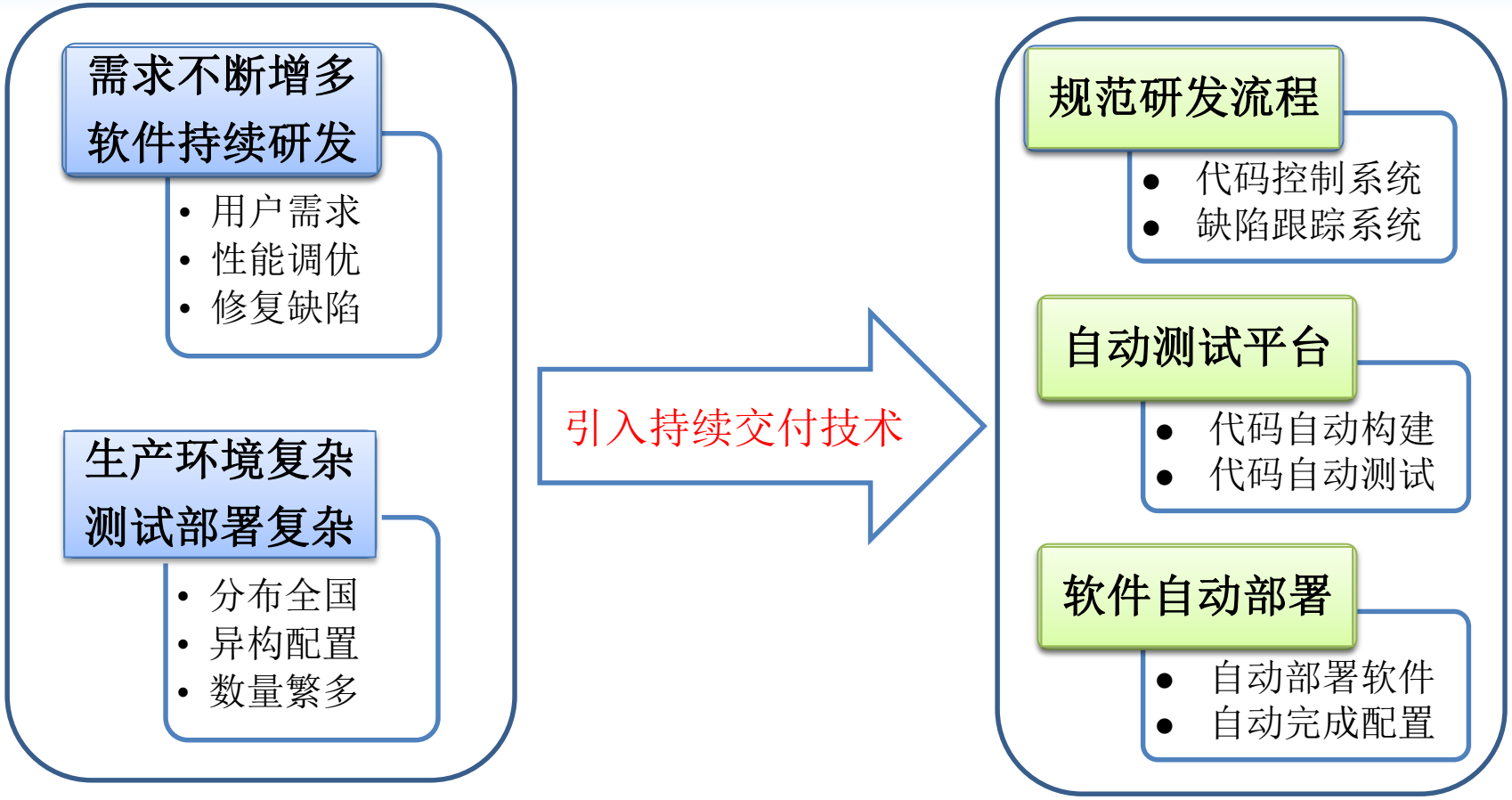
机时结算

为多种计费方式提供基础数据

- **分布式计算资源监控系统构建**
  - 适用于高性能计算环境的部署
  - 扩展插件：监控集群数据和环境服务的状态



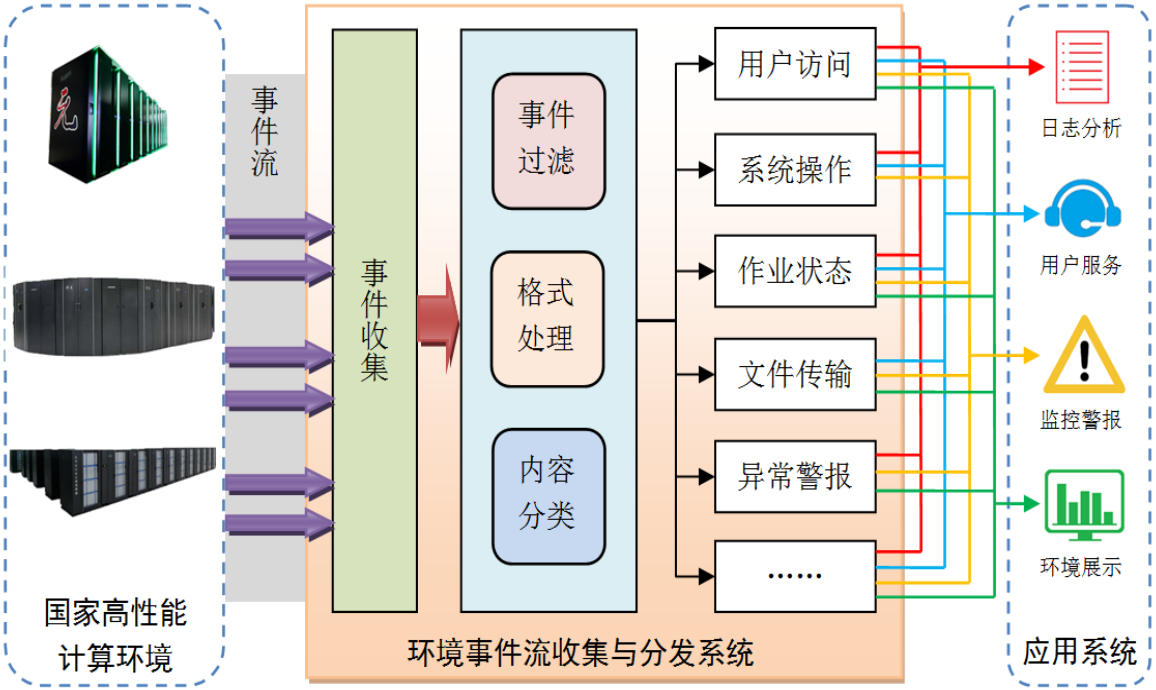
SCE软件快速部署

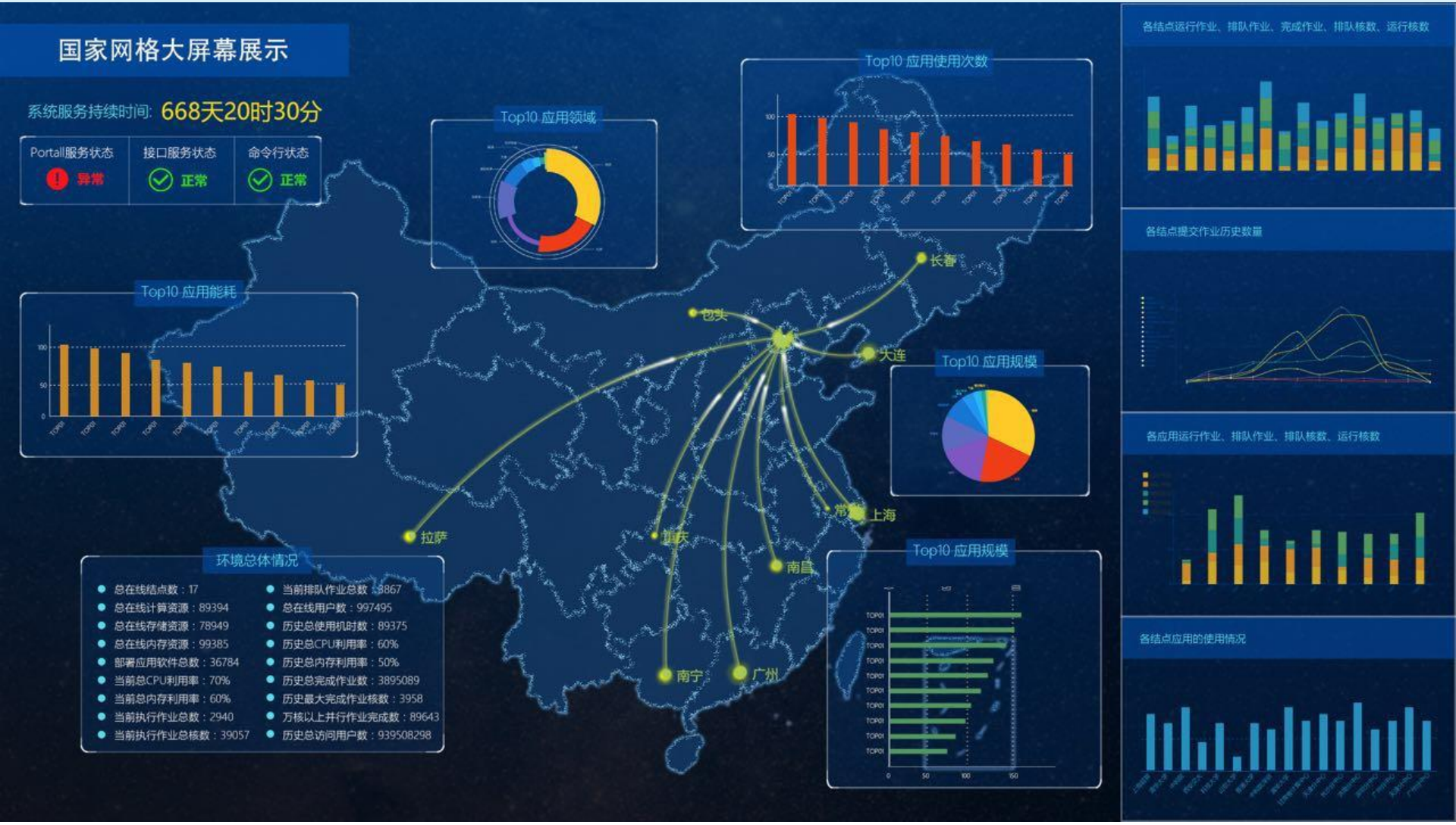


目标：缩短SCE交付周期，快速完成安装升级。

环境事件流处理与分发系统

- 针对环境运行时的各类事件
  - 收集、解码、过滤、处理、分类、分发
- 实时掌握环境运行情况
- 应用于环境展示、异常报警、用户推荐、行为分析等




国家网格大屏幕展示
系统服务持续时间: 668天20时30分
Portall服务状态
异常
接口服务状态
正常
命令行状态
正常
Top10 应用领域
Top10 应用使用次数
Top10 应用能耗
Top10 应用规模
Top10 应用规模
环境总体情况
总在线结点数：17
总在线计算资源：89394
总在线存储资源：78949
总在线内存资源：99385
部署应用软件总数：36784
当前总CPU利用率：70%
当前总内存利用率：60%
当前执行作业总数：2940
当前执行作业总核数：39057
当前排队作业总数：3867
总在线用户数：997495
历史总使用机时数：89375
历史总CPU利用率：60%
历史总内存利用率：50%
历史总完成作业数：3895089
历史最大完成作业核数：3958
万核以上并行作业完成数：89643
历史总访问用户数：939508298
长春
包头
大连
上海
拉萨
南昌
南宁
广州
各结点运行作业、排队作业、完成作业、排队核数、运行核数
各结点提交作业历史数量
各应用运行作业、排队作业、排队核数、运行核数
各结点应用的使用情况

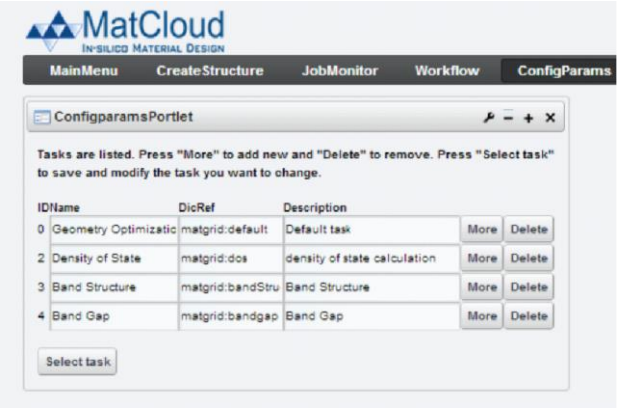
用户使用和技术支持情况

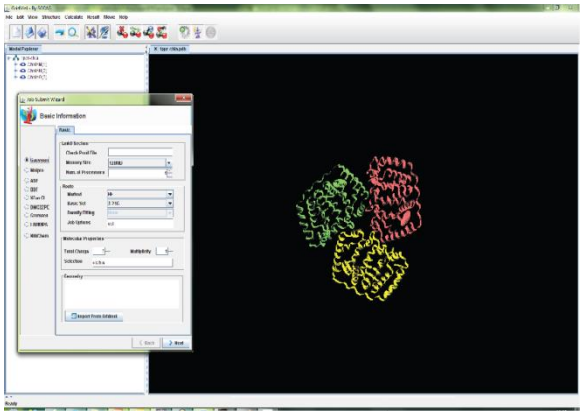
- 累计网格机时：1.99亿CPU小时
- 累计网格作业：875,036个

- 累计环境账号： 18,787个
- 累计技术支持邮件：5045封

数据统计时间：2018年5月

应用案例 – 社区建设

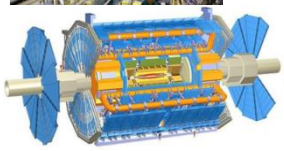


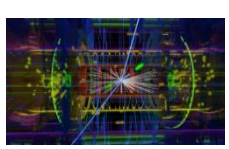


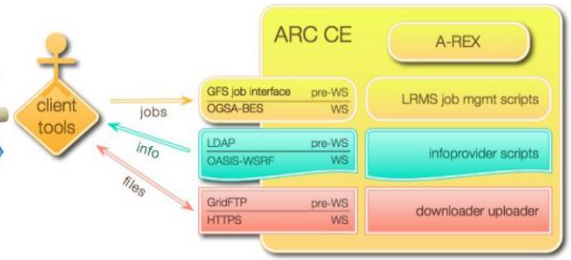
面向领域用户提供定制化、专业化的计算服务
已支持生物、化学、材料等10余个社区平台建设

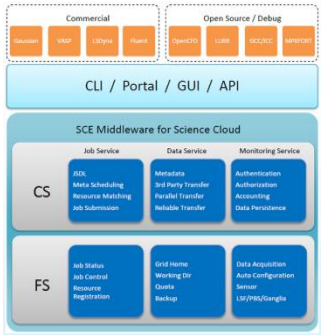
应用案例 - ATLAS计算

- 与CERN的ATLAS项目组合作
  - 使用SCEAPI与ARC-CE成功对接
  - 国家高性能计算环境为ATLAS提供计算服务
- 国际交流
  - 出访CERN，参加ATLAS站点大会并作报告
  - 参加ATLAS FR-Cloud讨论会并作报告
  - 参加CHEP2016大会并作报告



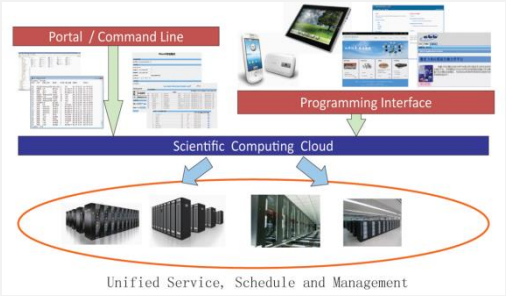






未来工作方向

优化核心软件结构

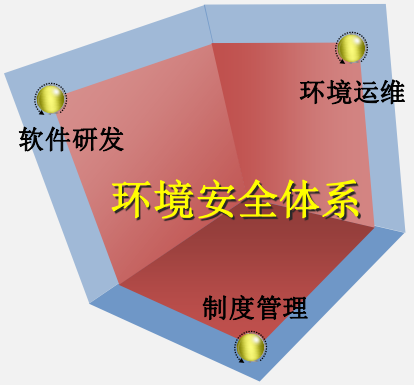


- 作业调度优化
- 微服务结构优化
- 持续软件部署
- 事件监控处理

建立多运行中心模式

- 分布式环境
- 高速网络互联
- 强容灾能力

构建国家高性能计算环境安全体系



- 账户认证与授权
- 资源和接口授权
- 数据访问安全
- 应用安全审查

谢谢！

扫码申请帐号