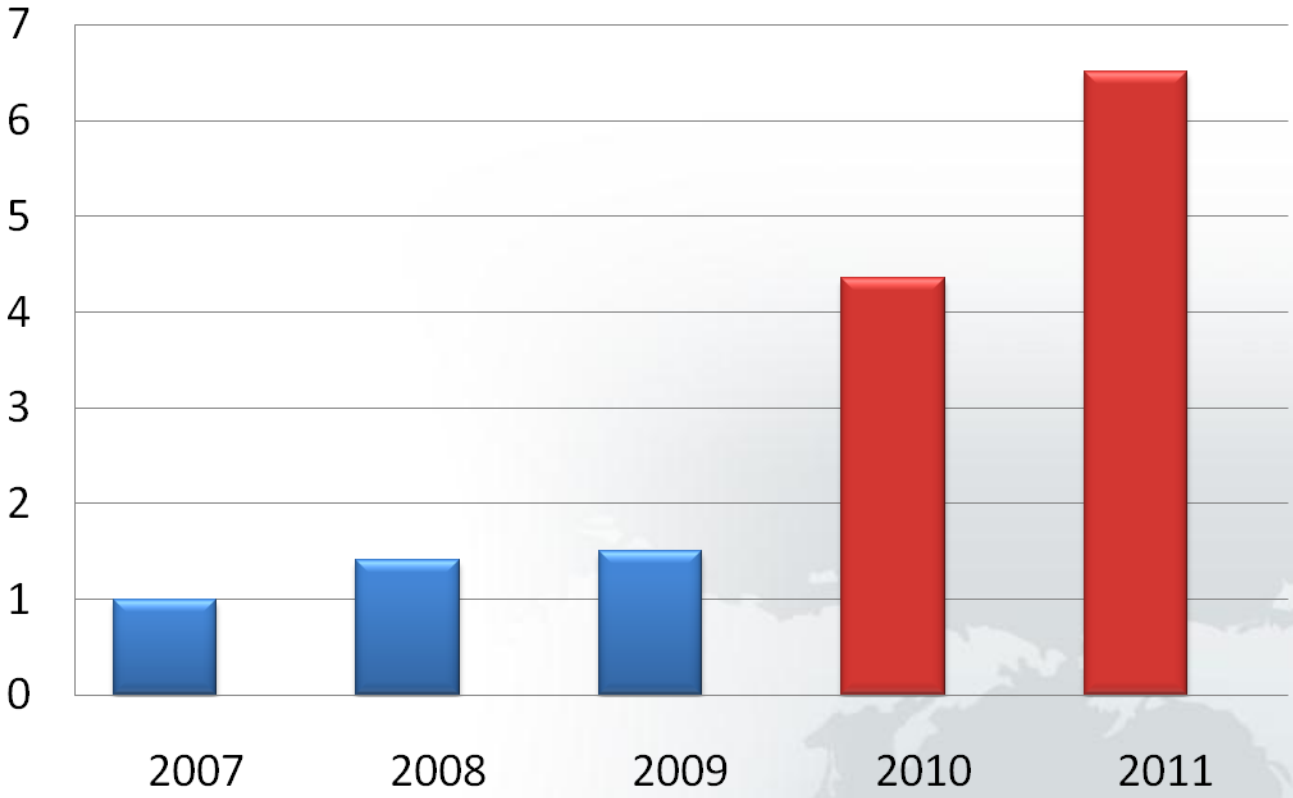
百度的脚印



今天的百度搜索离线系统

- ✓ 百亿网页需要在一天内完成分析
- ✓ 每天千亿级链接分析
- ✓ 数个异地机房
- ✓ 数十个异构集群
- ✓ 数万台机器

百度离线集群整合之道

Baidu 百度

内 容

- 背景
- 目标
- 架构
- 效果
- QA

目 标

“简单、可依赖”的CPU/MEM池

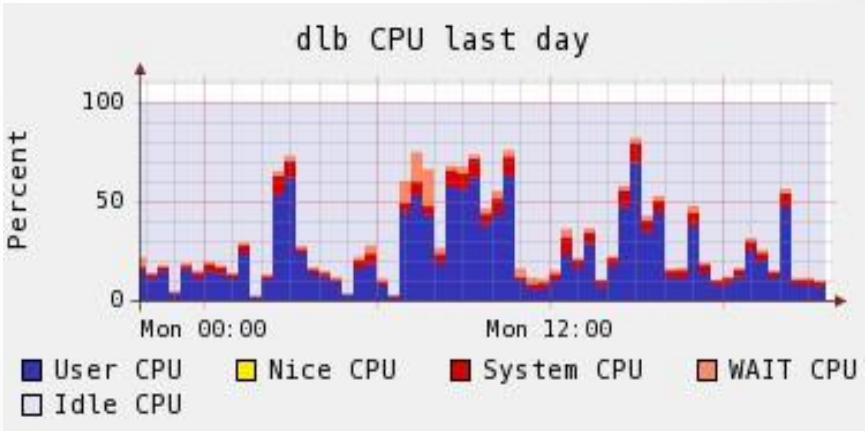
目标--- 整合的资源

System	硬件偏好	OS特性	分布式系统	使用规律	运维要求
网页库	IO吞吐	block access	自有系统	稳定	<1分钟
链接库	IO吞吐	block access	开源+自主系统	无规律	数分钟
基础检索	CPU,MEM,IO	mem,cache,tcp/ip,cgroup	无	稳定、有规律	<<秒
调研	MEM,IOPS	mem,cache,tcp/ip,cgroup	无	稳定	数分钟

目标--- CPU使用规律

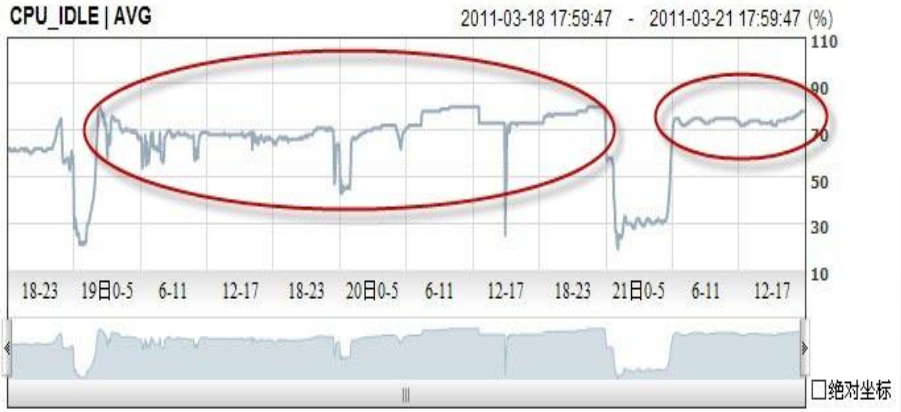

CPU_IDLE | AVG
2011-03-18 17:59:47 - 2011-03-21 17:59:47 (%)
110
90
70
50
30
10
18-23
19日0-5
6-11
12-17
18-23
20日0-5
6-11
12-17
18-23
21日0-5
6-11
12-17
绝对坐标


CPU_IDLE
2011-03-11 21:39:00 - 2011-03-14 21:39:00 (百分率)
100
80
60
40
20
0
12日0-5
6-11
12-17
18-23
13日0-5
6-11
12-17
18-23
14日0-5
6-11
12-17
绝对坐

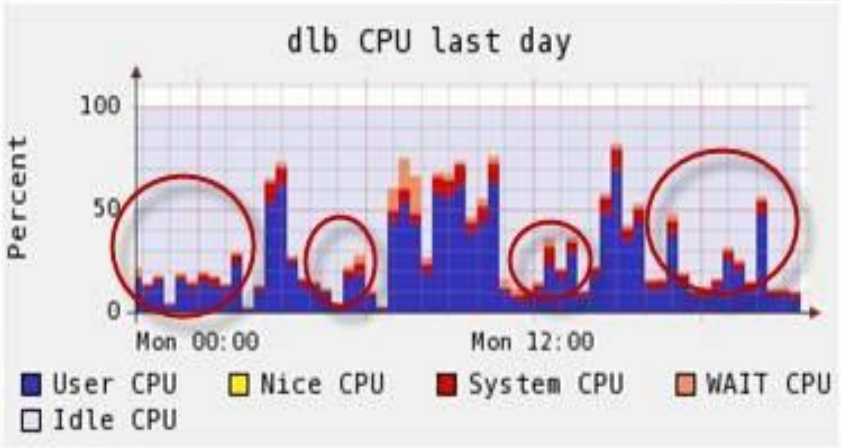

dlb CPU last day
Percent
100
50
0
Mon 00:00
Mon 12:00
User CPU
Nice CPU
System CPU
WAIT CPU
Idle CPU


CPU_IDLE
2011-03-13 21:41:05 - 2011-03-14 21:41:05 (百分率)
150
120
90
60
30
0
14日0-5
6-11
12-17
18-23
绝对坐

目标--- CPU使用规律


CPU_IDLE | AVG
2011-03-18 17:59:47 - 2011-03-21 17:59:47 (%)
110
90
70
50
30
10
18-23
19日0-5
6-11
12-17
18-23
20日0-5
6-11
12-17
18-23
21日0-5
6-11
12-17
绝对坐标


CPU_IDLE
2011-03-11 21:39:00 - 2011-03-14 21:39:00 (百分率)
100
80
60
40
20
0
12日0-5
6-11
12-17
18-23
13日0-5
6-11
12-17
18-23
14日0-5
6-11
12-17
绝对坐


dlb CPU last day
Percent
100
50
0
Mon 00:00
Mon 12:00
User CPU
Nice CPU
System CPU
WAIT CPU
Idle CPU


CPU_IDLE
2011-03-13 21:41:05 - 2011-03-14 21:41:05 (百分率)
150
120
90
60
30
0
14日0-5
6-11
12-17
18-23
绝对

Volunteer Computing system

project	start	Affiliation	Area	Peak#Hosts	Current status	Computing power
Bitcoin	2009	Bitcoin	Digital currengy	6,500 [40]	active	3500 TFLOPS[41]
Ibercivis	2008	National science agencies of Portugal and Spain	Biomedicine，physics，other	14,710[39]	active	6 TFLOPS
Rosetta@home	2005	University of Washington	Biology	100,000	active	100 TFLOPS
Einstein@home	2005	LIGO	astrophysics	280,000	active	370 TFLOPS
Climatepredictio n.net	2003	University of Oxford	Climate change	150,000	active	90 TFLOPS
BOINC	2002	University of California，Berkeley	Biomedicine，other	527,000	active	5430 TFLOPS
Foloding@home	2000	Stanford University	biology	406,000	active	7870 TFLOPS
SETI@home	1999	University of California，Berkeley	SETI	362,000	active	684 TFLOPS
GIMPS	1996	?	mathematics	10,000	active	40 TFLOPS
BVC	**2010**	**BAIDU**	**Search engine**	**12,000**	**active**	**740 TFLOPS(4th)**

ref : http://en.wikipedia.org/wiki/List_of_distributed_computing_projects

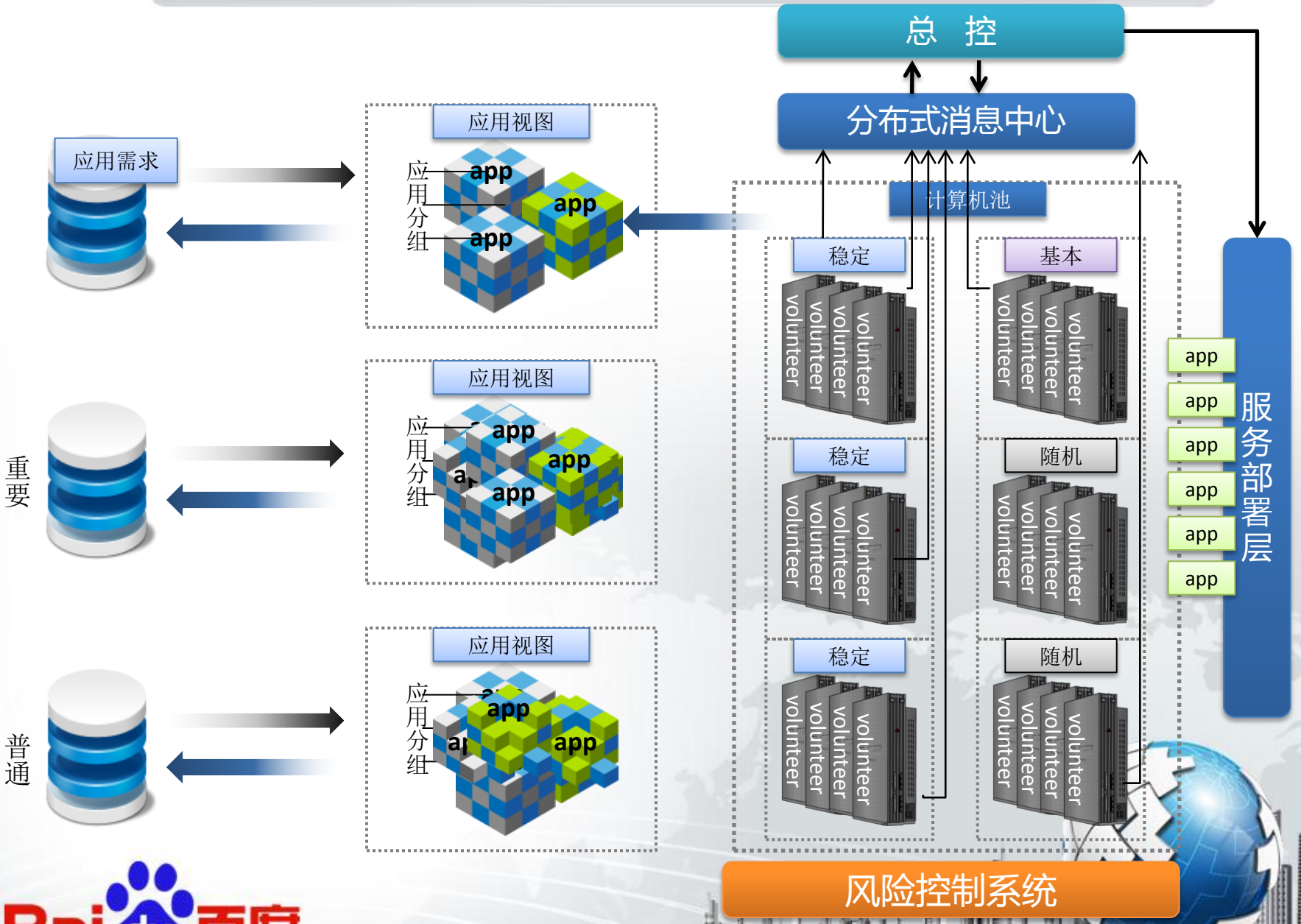
BVC : Baidu Volunteer Computing system

VC横向对比

- ✓ 复杂时效性需求
- ✓ 灾难处理的代价极大
- ✓ 网络整体的交互量大
- ✓ BVC不用解决的问题
  - 存储的整合
  - 计算可靠性

设计目标

- ✓ 优先级控制，全局负载均衡
- ✓ 风险控制， 最大利用率
- ✓ 插头、插座


总 控
分布式消息中心
应用需求
应用视图
应用分组
app
app
app
计算机池
稳定
基本
volunteer
稳定
随机
稳定
随机
app
app
app
app
app
服务部署层
重要
普通
风险控制系统
Bai du 百度

BVC总控

- ✓ 任务启停，权限控制，合法性校验
- ✓ 优先级控制
- ✓ 负载均衡
- ✓ BVC 总控不是系统中的单点

BVC优先级控制

- ✓ 时效性需求
  - 秒（分钟），小时，半天，天，周，月，季度
  - 优先级动态调整（时效性，相关性 ，覆盖率）
- ✓ 基于优先级分级控制
  - 优先级越高，越早调度
  - 同一级别内部基本均匀
  - 被延迟越久的任务，“级内优先级”越高
  - 优先级动态调整，人工确认

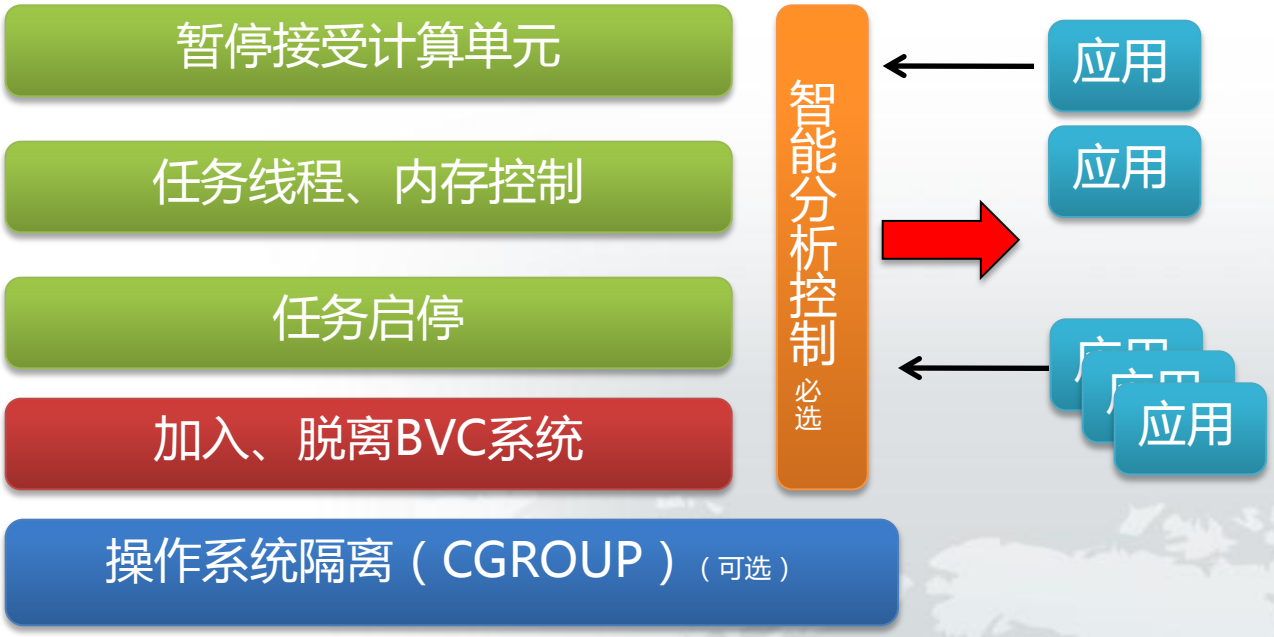
BVC负载均衡

- ✓ 通过volunteer分组，控制组间均衡
- ✓ 分组内部通过计算单元轮询控制均衡
- ✓ 任务在集群之间迁移

BVC 风险控制 最大利用率

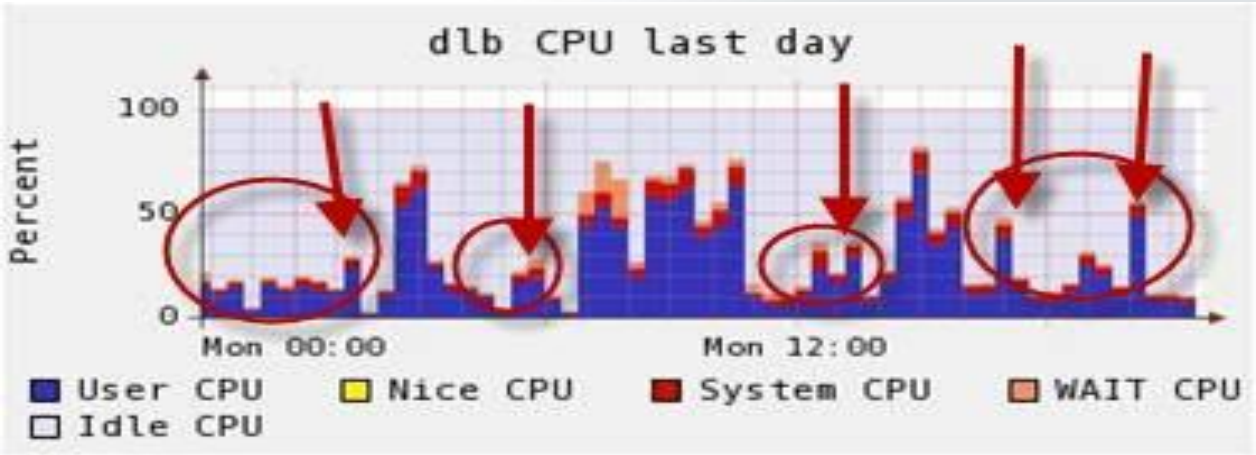
- ✓ 风险控制
  - 内核控制（ <<s ）
  - 智能分析 (10s, 30s)
- ✓ 最大利用率
  - 有规律：可配置，准确率很高
  - 无规律：行为分析，动态调整

BVC 风险控制体系


暂停接受计算单元
任务线程、内存控制
任务启停
加入、脱离BVC系统
操作系统隔离（CGROUP）（可选）
智能分析控制 必选
应用
应用
应用

BVC 风险控制 vs 利用率

暂停接受计算单元



BVC 风险控制 vs 利用率

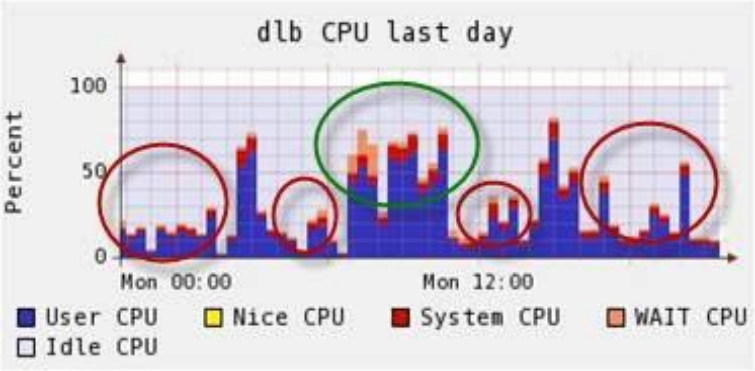
任务线程、内存控制

total mem = default + thr# * thr_mem

thr # = min (cpu_idle, free_mem)

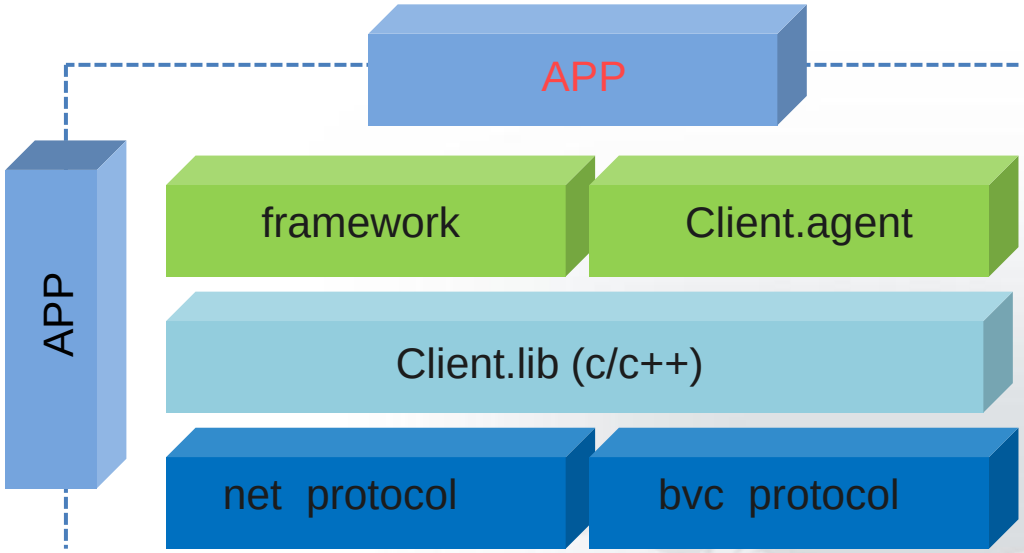
BVC 风险控制 vs 利用率

任务启停

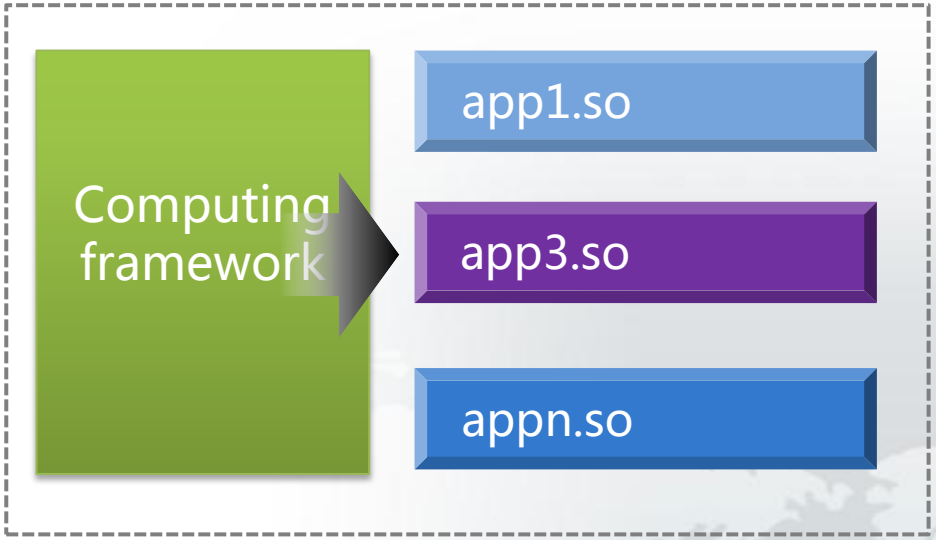


Baidu 百度

BVC 应用接入体系


APP
APP
framework
Client.agent
Client.lib (c/c++)
net protocol
bvc protocol

BVC 应用CF接入



BVC的实际应用

- ✓ 部分小规模的调研完全由BVC支持
- ✓ 主要业务之一的预算的 20%由BVC支撑
- ✓ 2011全年6月的计算能力峰值将达到1PFLOPS
- ✓ 2012：
  - 80%的网页搜索离线集群会加入bvc系统
  - 提供离线预算的30~40%

经验 教训

- ✓ 集群整合是成长过程中的痛苦
- ✓ Volunteer computing System是很好的选择
- ✓ 优先级控制，风险控制，简明易用的接口

Q&A

Thank you !

QCon

杭州站 · 2011年10月20日~22日

www.qconhangzhou.com（6月启动）

QCon北京站官方网站和资料下载

www.qconbeijing.com

全球企业开发大会

THE ANNUAL INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE