谷歌数据中心供电系统介绍

腾讯网络平台部数据中心规划组 李典林 曾宪龙
邮箱：dufresne545@gmail.com

2011 年中，谷歌公布了其在 2010 年全年的数据中心耗电情况。根据谷歌提供的数据，这家互联网公司一年的电力消耗量高达近 23 亿千瓦时，比 21 万个美国家庭一年的用电量加一块儿还要多。据斯坦福大学的 Jonathan Koomey 估计，截至 2010 年底，谷歌共拥有近 90 万台服务器，约占全球 3％的服务器数量，但只使用全球数据中心 1％的电力，目前其部分数据中心的年均 PUE 约为 1.12 左右，显然谷歌数据中心的运作比其他数据中心更为高效。我们知道除了制冷系统外，数据中心的供电系统也是能耗很大的一块，因此有必要研究一下其供电系统，虽然不一定适合中国国情和一般用户，但其理念和思路却是非常值得借鉴的。

图 1 是谷歌某个数据中心外部供电照片，由于谷歌的数据中心体量通常都很大，比如达到 30-40M 以上的用电，因此往往机房周边专门区域建设有专用的变电站给庞大的机房供电。

图 1 谷歌数据中心外部供电照片

由于篇幅有限，变电站及中压部分就不再展开，但其总体思路是采用中压配电输送到机房周边，靠近负载就近经变压器降压成低压，再通过低压母线输电到机房内的 IT 机柜上。从图 1 我们可以看到模块化的户外型变压器及低压配电柜环绕机房周边，采用集装箱型的柴油发电机组作为变配电的供电投切备份，柴发风管直立到屋顶上排风。经过变压器变压后的市电通过母线槽或者线缆直连到机房内的机柜上方，直接给自带分布式 UPS 的服务器供电。

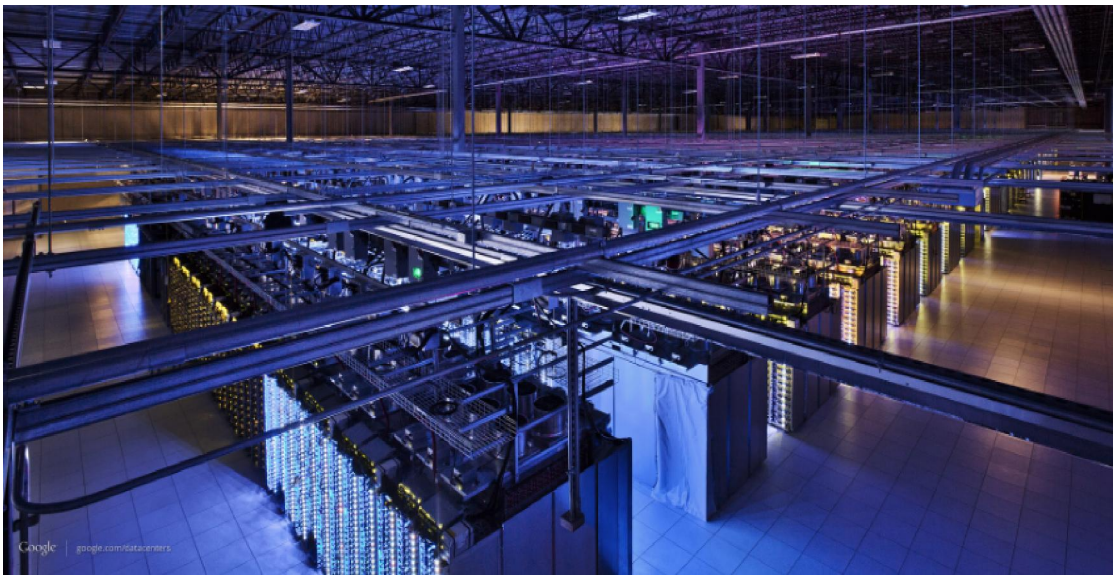


图 2 谷歌数据中心内部供电示意

图 2 则是谷歌数据中心内部的俯视图，从这个照片我们可以看到市电经前面描述的室外变压器降压后直接通过母线槽的方式架设在每排机柜顶部，再用机柜顶部的配线盒连接到每个机柜 PDU，具体细节我们后面会介绍。由于谷歌定制的服务器上自带有分布式小 UPS，因此谷歌的数据中心内部不再有 UPS 室和电池室等，也没有列头柜等二次配电环节，每个服务器直接采用市电直供技术，达到接近 99.9%的供电效率。如图 3（a）所示的每个机柜直接从机柜顶部母线槽上安装的配线盒取电，这种供电架构非常简单清爽，大大减少了线缆的采购和工程施工，而且非常灵活便于机柜扩充和带电检修维护等，运营起来也非常简单，还可以根据机柜的功率和用电可靠性情况灵活调整配线盒来满足不同设备的供电需求。此外，变压器、中低压以及柴发供电外置，且无集中 UPS 和电池等，机房内的空间利用率也非常高。

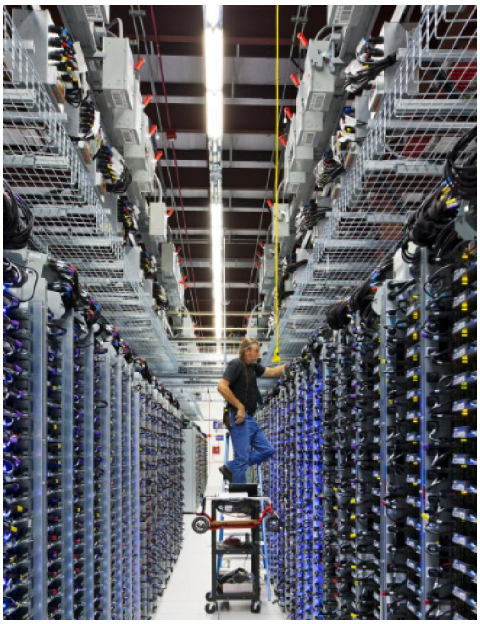


图 3（a）机柜顶部的母线及配线柜

图 3（b）谷歌的自带 UPS 服务器

如图 3（b）所示谷歌的带小 UPS 的市电直供服务器大家都应该非常了解了，这里不再详述，只简做要介绍。其原理是在服务器内安装底部的 12V 黑色铅酸电池用于市电停电保护，市电正常时候，由于没有外部 UPS，市电直接给服务器供电，达到 99.9%的供电效率；当市电停电后，直接挂接在 12V 输出上的电池短时放电，直至室外的柴发启动恢复服务器电源带载。电池参与放电的时间基本不到一两分钟，因此电池的容量很小，大约只有 3.2Ah，备电时间远远小于传统数据中心 15 到 30 分钟的电池备电需求，因此对柴发的启动要求很高。

我们前面知道谷歌的柴发是模块化直接安装在变压器旁边的，很有可能是低压柴发，其启动很快。而且每台柴发对应一个变压器，没有复杂的柴发并机以及启动时序等问题，因此正常情况下柴发启动时间可以控制在十几秒以内，一两分钟的电池备电时间基本上是够了，但这对运维水平要求就非常高了。当然谷歌的软件架构和业务备份方面也足够强壮，甚至部分设备停电也不会影响到业务正常运行，因此只有强大的技术实力才可以采用这种供电架构。

图 4 是谷歌三联柜 102a-c 的正视图，以及顶部配电系统 100，后者不仅给三联柜服务器供电，还给相关的空调等支撑设备供电，比如机柜顶部的风扇等。三联柜底部装有滚轮 105，用于支撑机柜，并且方便现场运营人员灵活搬运，每个机柜高度约为 30 个导轨，估计每个设备高度约 1.5U，总计高度 45U 左右，整个三联柜总共有 90 个设备安装空间。这种设计使得整个三联柜不会太高，无需凳子或者梯子等来协助。三联柜横向宽度方面也不会太宽，叉车等工具可以简易搬迁，甚至现场安装人员可以徒手在数据中心内很容易搬动到需要的地方。

空气循环单元 107 会被安装在背靠背的两组三联柜之间，该单元包含两组三联柜之间的锁固框架，以及一到两个的空调盘管，还有热通道顶部的一些风扇等。三联柜排出的热空气

可以进入框架内，然后被空调盘管制冷，最后从顶部的风扇被吹回整个机房冷环境中。空气循环单元 107 的框架上还架设有支撑架 108，支撑架 108 用于安装线槽 110 和 112 等，分别用于布设供电线缆和光纤或铜缆等。空气循环单元上还有一些网线等，用于控制风扇的转速，并将收集到的压力、温度、设备故障信息等送回到机房监控中。

三联柜顶部的配电系统 100 则包括 IT 机柜供电的配线盒 116，该配线盒向上可通过 118 接入口从一根或者两根供电母排上获得电力，向下则通过配电空开 122 以及线缆 120 等给 102a-c 的三联柜每个机柜分别供电。比如每个机柜有三根线 120 到机柜 PDU，那么 122 配电空开就总共有 9 个微断连接到 9 条输出线缆 120 上，或者 3 个 3P 的空开。

由于三联柜通常都是工厂预制并安装好 TOR 的整机柜，当三联柜被滚动安装锁固在空气循环框架 107 后，就可以直接从线槽 110 或者 112 上连接上联光纤。然后供电的配线盒 116 向上则从供电母线取电，向下则给三联柜机柜的 PDU 供电。最后等所有的线缆都连接好并带电后，三联柜内的服务器上电开始加载系统开始工作。如前面所述，每个机柜有 30 个导轨安装 30 台服务器，每个机柜采用三相空开，每相空开带载 10 台设备，每组三联柜共有 9 路输出空开，如果任何部分发生故障，那么最多也只会影响到 1/9 的设备，在保证人身安全的同时，尽量缩小故障影响面。

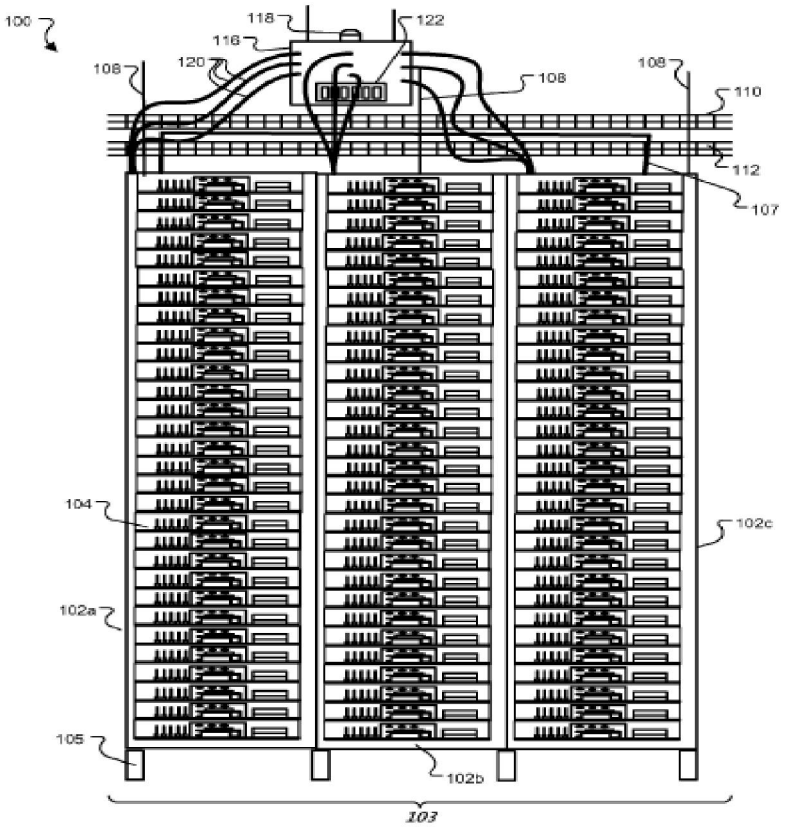
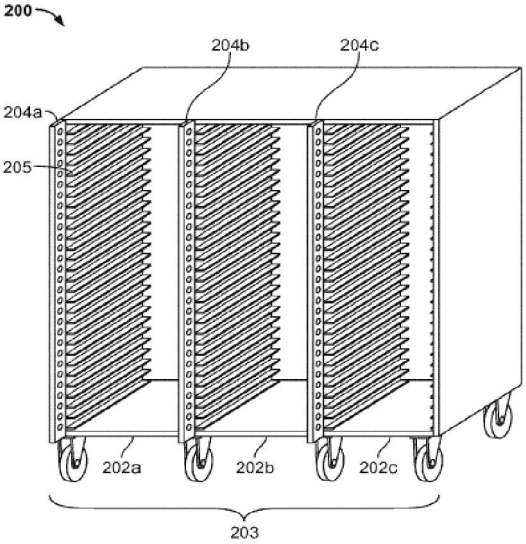


图 4 谷歌三联柜机柜及其顶部配线盒

图 5（a）是谷歌三联柜更为详细的展示，三联柜 202a-c 内的三个机柜整个连为一体，导轨 205 用于支撑服务器的安装。标识为 204a-c 为机柜上的 PDU 垂直安装在机柜前面的左侧，PDU 上有 30 个左右的母头插座用于安装服务器电源插头。由于 PDU 和服务器非常靠近，而且插座基本平齐于服务器前面板，因此可以采用非常短的电源线用于连接服务器和 PDU，电源线缆安装美观且管理方便。图 5（b）是安装了服务器的三联柜，这些服务器可以是计算、存储等各种类型的 IT 设备。整个三联柜通过底部的滚轮方便搬运和安装，并且这些三联柜以及要安装的服务器会以整机柜的方式在工厂整体安装调试好，搬运到现场后整体交付，即插即用，满足快速部署和建设的需求。



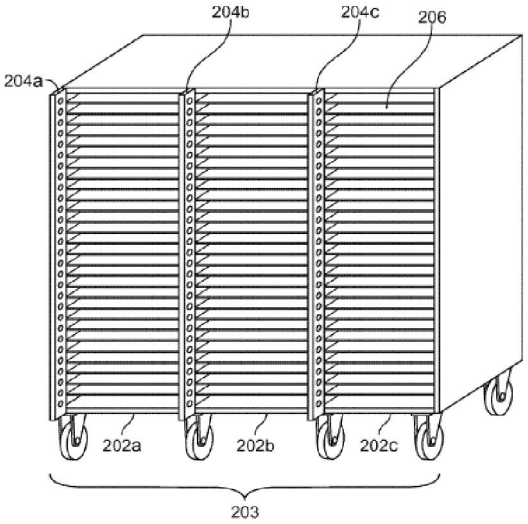
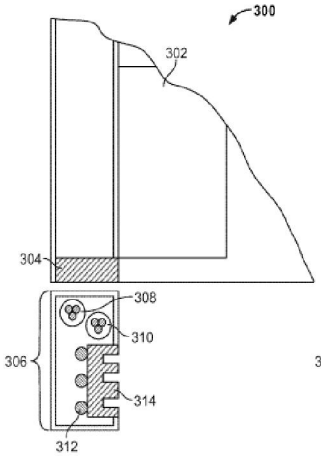
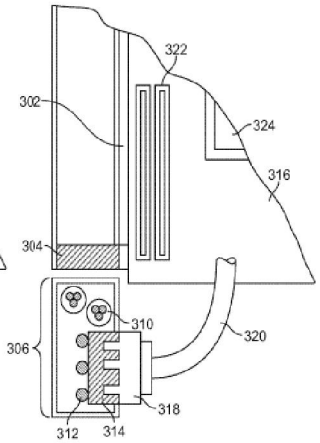


图 5（a）三联柜框架示意　　图 5（b）装满服务器的三联柜机柜

图 6（a）是 PDU 及机柜一角的俯视图，其中 302 是用于支撑待安装服务器的导轨，而 304 则是 PDU 的支撑架，用于固定垂直的 PDU 306 到机柜上。PDU 内部包含三根导线 308、310 和 312，每根导线连接 PDU 上 1/3 的插座，比如导线 312 连接到 PDU 顶部的 10 个插座 314，310 负责中间的 10 个插座，而 308 则负责底部的 10 个插座。图 6（b）则描述了服务器主板 316 被安装在导轨 302 上，以及和 PDU 的电源线连接示意。服务器 316 通过电源线 320 以及电源插头 318 从 PDU 上取电，当然这里电源线只是示意，实际上谷歌的服务器电源往往安装在热通道侧，电源线穿过服务器到机柜前面的冷通道来取电，满足谷歌服务器全正面冷通道维护的需求。图 6（c）更为详细的展示了导轨 302 和 PDU 306 之间的位置，插座 314 和导轨 302 一一对应逐个匹配，所以服务器电源线可以就近很好得连结到每个插座，当然如果有需要，也可能会有冗余的插座用于安装其他需要的设备，比如交换机等。





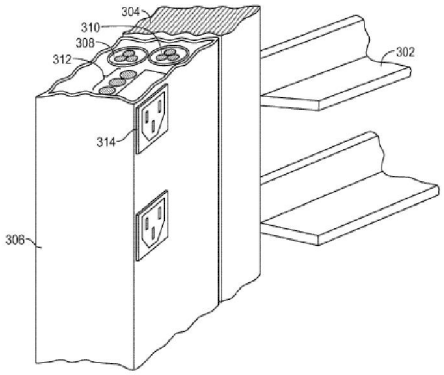


图 6（a）PDU 俯视图　　图 6（b）服务器安装示意　　图 6（c）PDU 及导轨正视图

图 7 是 PDU 的示意图，整条 PDU 分为 6 个子部分，从 402 到 412，每个子部分则包含 5 个插座。当然也可以分成其他数量的子部分和不同的插座数，但每个子部分之间是没有直接电气连接的，且每个部分会设置有独立的空开或者熔丝等来保护该子部分。标识为 418、420 和 422 的三个插座是三个电气连接口，用于整条 PDU 从低压母线排上的配线盒来取电。比如前面的例子，418 输入的电源线会覆盖到 2 个 PDU 子部分，共连接 10 个插座。同样的，420 和 422 输入也各覆盖 10 个插座。PDU 上还安装有部分的熔丝保护槽 414，用于安装和

快速更换熔丝。比如，图 7 中 414 熔丝槽内的熔丝可以用于保护和隔离 406 部分的 5 个插座。如果 406 子部分 5 个插座的总负载电流不超过 15A，那么 418 的输入线缆会选择载流量为 30A 的电源线，整条 PDU 则可以承载 3*30A 的 IT 负荷。在 230Vac 到 240Vac 供电情况下，418 输入会选择 10AWG 或者 4mm^2 的线缆。根据谷歌服务器 300 到 400 瓦的典型功率，单机柜功率约在 9-12KW 附近。此外，上级配线盒内会选择容量较大，而且动作响应时间较慢的开关，这种情况下靠近负载的熔丝会在故障发生时先于上级的开关动作，快速切断负载，并且确保不会影响到上级开关，这种上下级开关和熔丝的选择性问题会在后面详细描述。

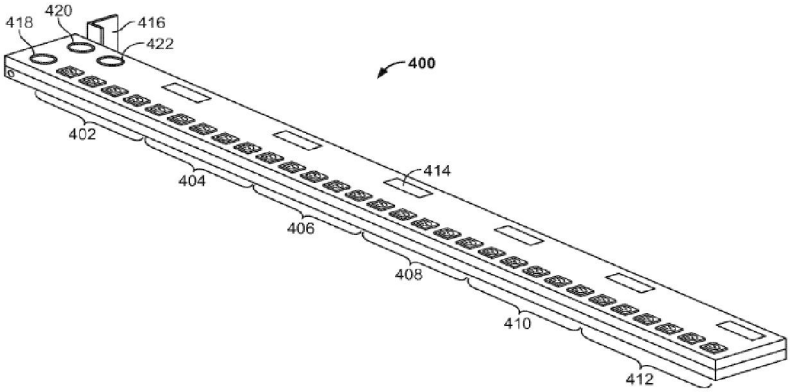


图 7 谷歌 PDU 的示意图

图 8 是标识为 600 的 PDU 内部电路图，包含三根导线 602、604 和 606，每根导线及其分支都会连接到共两根熔丝上，每根熔丝又覆盖 5 个插座，这样每根导线覆盖 10 个插座。每根熔丝的载流量也不能选择过大，以至于不能保护其上级连接的线缆和空开等。比如导线 602 的上级空开为 30A 的微断，那么熔丝就必须选择小于 15A 的，这样确保两根熔丝的载流量不会超过上级空开 30A 的分断能力。同时选择快熔型熔丝，以保证机柜 PDU 内配置的熔丝动作时间小于配电箱内上级微型断路器的动作时间。5 个插座对应 1 个熔丝，熔丝槽设计便于熔丝维护更换，以减少修复时间。采用这种一根熔丝只带 5 个插座的设计，保证了任意一台服务器的电源短路故障只会影响到该子部分的 5 台服务器，避免该短路故障扩散到整个机架。同时熔丝可在 PDU 的熔丝槽内快速更换，也减少了被影响服务器的故障修复时间。

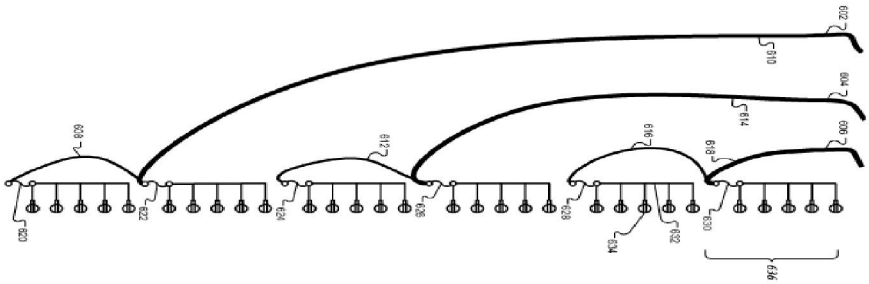
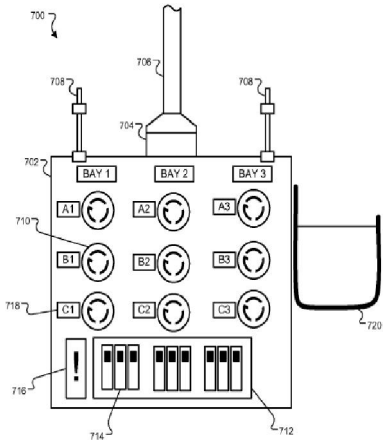
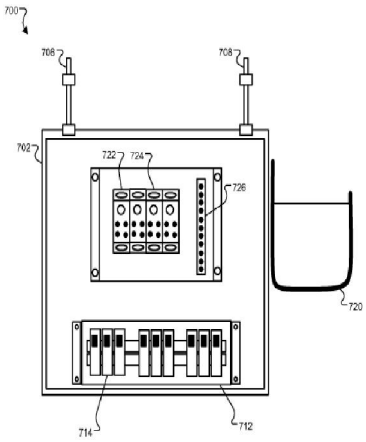


图 8 谷歌 PDU 内部电路图

图 9（a）展示了带 9P 空开的配线盒正视图，输入线缆 706 可从母排或者配电柜取电，其线径可选择 AWG 2 号线或者载流量达到 100A 的线缆。配线盒面板上有 9 个型号为 Leviton L6-30R 的输出母头和 9P 控制空开。配线盒侧面还有个线缆捆扎钩 720，用于输出线缆的捆扎整理，线缆美观的同时还能减少下垂线缆的重量对配线盒的拉伸应力。图 9（b）则展示了配线盒内部的配线排细节，比如输入配线排将输入的三根火线分配到 9 个空开上，还有中线排和地线排等用于 9 个支路的输出，并且接地排同时还会连接到配线盒外壳上作为接地保护。图 9（c）则是配线盒的侧视图，展示了输出线缆如何挂接在线缆钩上，减少线缆重量对线缆输出接头的拉伸应力。这种配线盒设计有效节省了机房空间，并满足快速部署需求。





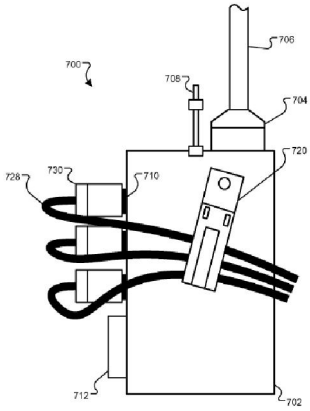


图 9（a）配线盒正视图　　图 9（b） 配线盒内部　　图 9（c） 配线盒侧面理线钩

图 10 是四组三联柜以及顶部配电盒的示意图，每组三联柜 1002 从机柜顶部带 9 路输出的 IT 配线盒 1006 上取电。除此之外，三联柜顶部还有另外一个带 12 路输出的空调配线盒 1010，用于给背靠背两排机柜间热通道顶部的置顶空调风机供电，后面会有详细介绍。

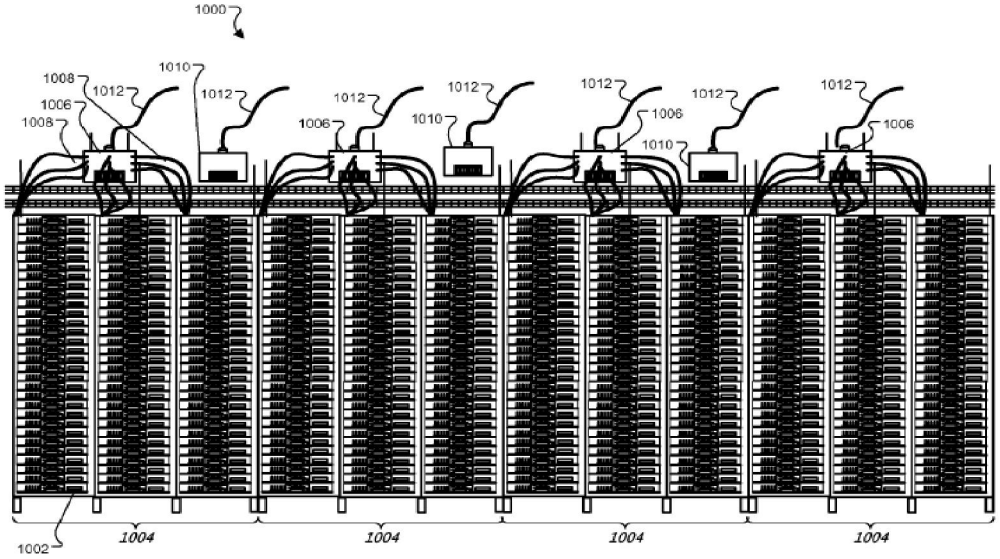


图 10 多组三联柜及配电正视图

图 11 是前面图 10 中四组三联柜及其配电部分的俯视图，同样的，带 9 路输出的 IT 配线盒 1106 给三联柜 1102 供电，而带 12 路输出的空调配线盒 1110 则是给热通道顶部置顶空调的 3 个大风扇供电。不管是 1106 的 IT 机柜供电配线盒还是 1110 的空调风扇供电配线盒都是通过供电线缆从两排机柜中间顶部的供电母排 114 上取电。

置顶空调 1118 包含了标识为 1120 的 3 个散热风扇和标识为 1122 的变频驱动控制器，后者用于控制风扇的转速，同时将置顶空调的故障信息上传到机房监控中。带 12 路输出的空调配线盒 1110 中的 9 路输出用于给三个大风扇供电，而剩余的 3 路输出则给风扇驱动控制器供电。而置顶空调 1118 的 6 个风扇及两个控制器分别从两套 12 路输出空调配线盒 1110 来供电，每边各一个，避免某个配线盒故障导致整个空调停机，来提高空调系统的可靠性。

置顶空调 1118 的宽度或者相邻空调之间的距离和三联柜 1104 的宽度可能会不一样，或者说下图的横向方向上置顶空调的数量和三联柜的数量可以一样，也可能不一样，这种配置会随着三联柜功率的不同灵活调整。比如三联柜 1104 放置高功率服务器情况下，置顶空调和三联柜会一一对应逐个排开，但如果三联柜 1104 改上架低功率的服务器后，可能会少安装一些置顶空调 1118，这些空缺出来的空调位置可以通过通道顶部的一整块铁皮来封堵。

由于整排机柜共享热通道，也共享冷通道，那么置顶空调空缺出来位置的热空气会横向流动或者上下流动，依靠旁边的置顶空调风扇来散热。还有整个机房大环境是冷通道，那么弥散在整个房间内的冷气可以较为均匀地被全部服务器风扇吸入，很少出现局部热点。

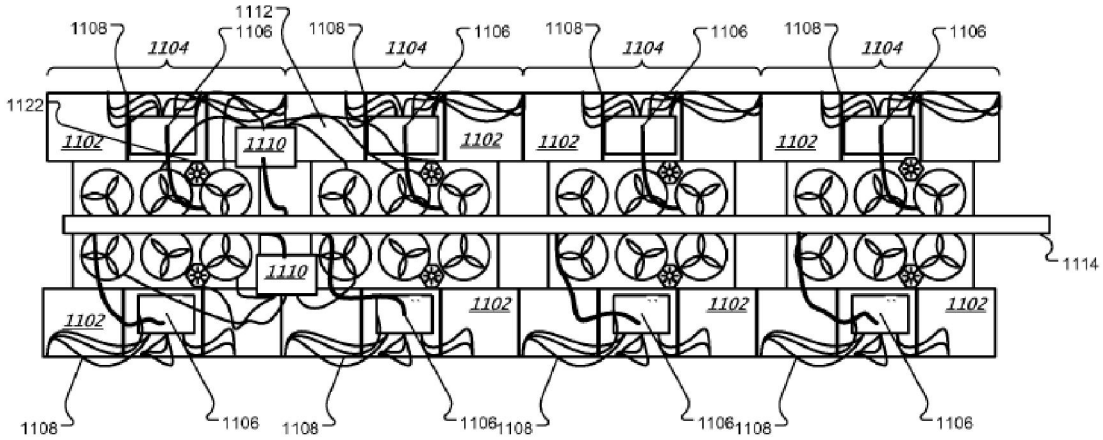


图 11 多组三联柜及配电俯视图

图 12 是数据中心机柜、空调和配电部分的剖面图，整个机房的底部是架空地板 1202 和顶部的天花 1204，整个机房的核心部分由两排机柜 1206a 和 1206b 以及置顶空调 1208、9 路输出的配线盒 1210a 和 1210b、12 路输出的配线盒 1212 以及母线排 1214 构成。两排机柜 1206a 和 1206b 从 9 路输出的配线盒 1210a 和 1210b 来取电，两排机柜后面热通道顶部则为置顶空调 1208，其包括 6 个轴流风机 1218、一对风扇控制器 1220 以及一对散热盘管 1222。每组风扇和控制器从带 12 路输出的配线盒 1212 上取电。顶部的母线排可能的载流量约为 1000A，覆盖多组三联柜机柜和置顶空调单元。

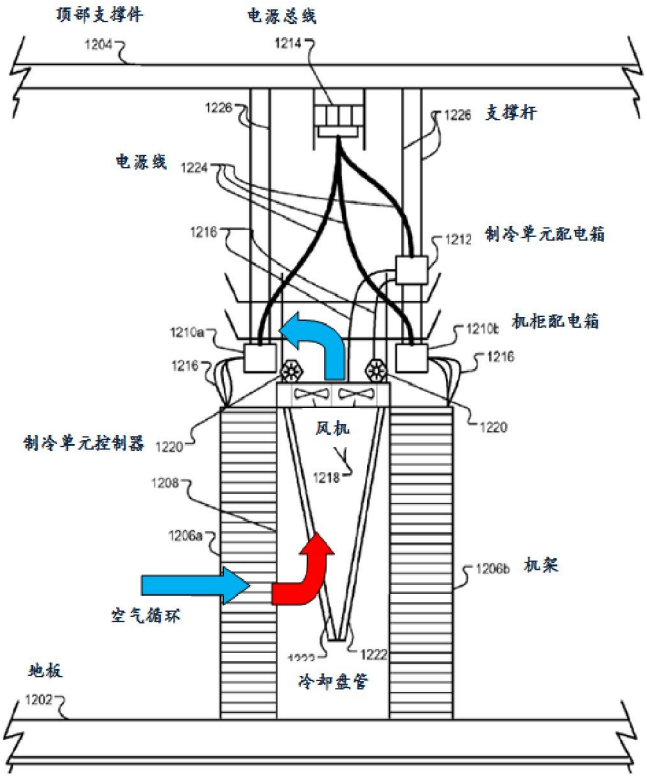


图 12（a）热通道剖面图

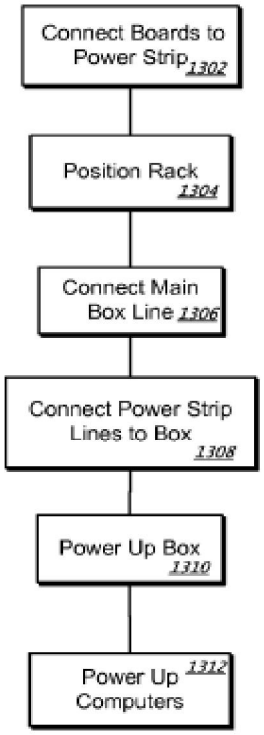


图 12（b）三联柜安装及上电

正常工作情况下，1206a-b 的服务器散发出来的热量进入共享的热通道内，经过盘管

1222 制冷后，被热通道顶部的风机 1218 吹回到整个机房大冷环境内，再重新被服务器风扇吸入，不断循环，机房大冷池环境如图 13（a）所示。或者如果刚好部分机柜顶部没有风扇或者该风扇出现故障，则该机柜散发出来的热量会横向流动到相邻带风扇的空调（比如进入或者流出此正视图）被相邻的空调给制冷，共享热通道内部如图 13（b）所示，这种冷通道和热通道共享的机制可以大大减少制冷系统的局部故障，提高系统可靠性。

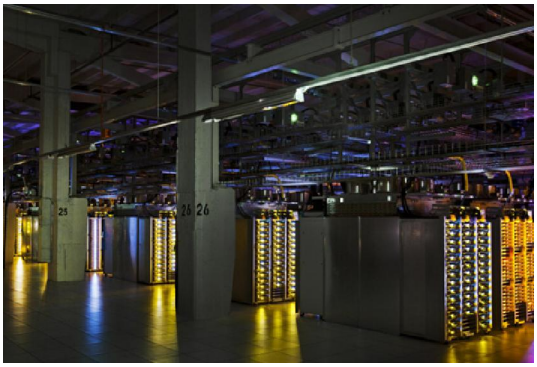


图 13（a）整个机房作为冷通道

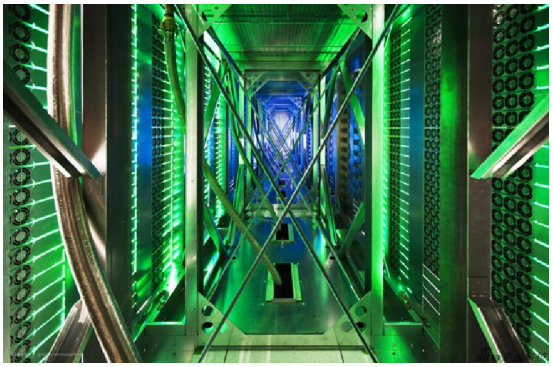


图 13（b）共享热通道内部的照片

图 12（b）是实际部署机柜上电的流程图，首先 1302 步骤在工厂将每台服务器上架并连好电源线到 PDU 上，然后是 1304 步骤将整机柜运送到机房现场泊位，在对应位置固定安装好整机柜并连好机柜接地线，接着是 1306 步骤在供电母线上安装 IT 机柜供电配线盒，并在 1308 步骤上连接每个机柜 PDU 和顶部配线盒，于 1310 步骤给配线盒总闸合闸上电，最后在 1312 步骤给每个机柜的 PDU 供电开启机柜服务器。我们之前在文章《谷歌的服务器内置 UPS 技术介绍》中了解到其服务器电源的内部作了上电随机延迟设计，因此不会在整机柜合闸的瞬间导致大电流冲垮熔丝或者支路空开，或者也可以在远程后台逐个上电，避免同时上电的启动冲击。

综上，本文简单介绍了谷歌数据中心的供电架构，有些盲人摸象，泛泛介绍了些内容，很多技术细节还有待进一步充实。但从我们了解的有限知识内，这个架构是对传统数据中心供电架构的颠覆，摒除了传统上依靠 UPS 等硬件冗余的低效率模式，追求能效无极限。同时还在高效建设、便捷运营以及系统可靠性等方面也有很多亮点。虽然这种完全定制的市电直供架构不一定适合国内数据中心用户，但其低成本、简单、高效的模块化建设思路非常值得国内学习。