

方正证券研究所证券研究报告

行业专题报告

行业研究

半导体行业

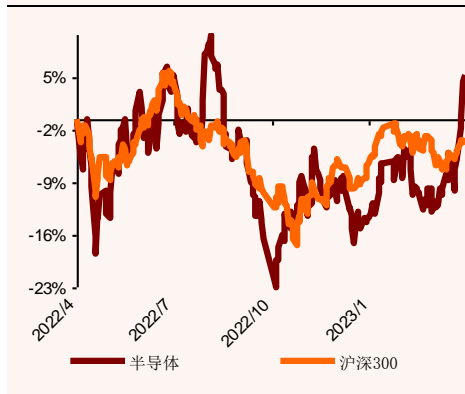
2023.04.08/推荐

分析师： 吴文吉
登记编号： S1220521120003

重要数据：

上市公司总家数	114
总股本(亿股)	790.24
销售收入(亿元)	3670.74
利润总额(亿元)	642.49
行业平均 PE	-102.17
平均股价(元)	75.72

行业相对指数表现：



数据来源：wind 方正证券研究所

相关研究

《AI 赋能智能家居》2023.04.06
《聚辰股份：加强存储战略布局，添砖加瓦助发展》2023.04.06
《AI 高算力催涨光模块 CPO 需求》2023.04.05
《AI 爆发，服务器模拟 IC 鹊起》2023.04.03

AIGC 算力大时代下，GPU 支撑强大的算力需求。ChatGPT 这样的生成式 AI 不仅需要千亿级的大模型，同时还需要有庞大的算力基础。训练 AI 现在主要依赖 NVIDIA 的 AI 加速卡，达到 ChatGPT 这种级别的至少需要 1 万张 A100 加速卡，而一颗英伟达顶级 GPU 单价高达 8 万元。

存算一体化突破算力瓶颈，GPU 封装进入正当时。在 AI 运算中，神经网络参数（权重、偏差、超参数和其他）需要存储在内存中，常规存储器与处理器之间的数据搬运速度慢，成为运算速度提升的瓶颈，且将数据搬运的功耗高。在算力芯片性能暴增的时代下，相关的封装产业链也逐渐的进入高速发展时期。

Chiplet 为主要发展方向，CoWoS 被广泛应用于 GPU 封装。Chiplet 将大型单片芯片划分为一组具有单独功能的小芯片单元，再通过跨芯片互联和封装集成，这就要求发展先进封装技术，提高布线密度和信号传输质量。CoWoS 是由台积电主导，基于中介层实现的 2.5D 先进封装，能达到封装体积小、功耗低、引脚少的效果。英伟达 V100/A100/H100 高端 GPU、AMD 新 MI 系列数据中心加速器芯片，均采用台积电 CoWoS 封装。

“GPU+存储器”的 HBM 封装模式突破了内存容量与带宽瓶颈。凭借 TSV 方式，HBM 将多个 DDR 芯片堆叠在一起，实现大容量，高位宽的 DDR 组合阵列，使 DRAM 从传统 2D 转变为立体 3D，比 GDDR5 节省了 94% 的表面积，充分利用空间，实现集成化。同时，HBM 大幅提高了容量和数据传输速率，具有更高带宽、更多 I/O 数量、更低功耗，革命性地提升了 DRAM 的性能。

建议关注：长电科技、通富微电、华天科技、甬矽电子、晶方科技。

风险提示：

- 1、AIGC 整体发展不及预期；
- 2、先进封装产能不足；
- 3、供应链波动，封装材料紧缺；
- 4、CoWoS 研发迭代进度不及预期；
- 5、国际贸易摩擦的风险。

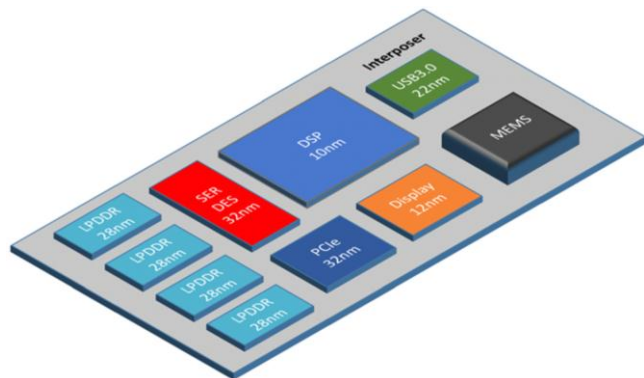
1 GPU 封装：大算力时代下，被寄予厚望的 Chiplet

AIGC 算力大时代下，GPU 支撑强大的算力需求。GPU 即图形处理器（英语：graphics processing unit），又称显示核心、视觉处理器、显示芯片，可以兼容训练和推理，被广泛运用于人工智能等领域。作为 AI 硬件的心脏，GPU 的市场被英伟达和 AMD 等海外巨头垄断。ChatGPT 这样的生成式 AI 不仅需要千亿级的大模型，同时还需要有庞大的算力基础。训练 AI 现在主要依赖 NVIDIA 的 AI 加速卡，达到 ChatGPT 这种级别的至少需要 1 万张 A100 加速卡，而一颗英伟达顶级 GPU 单价高达 8 万元。

存算一体化突破算力瓶颈，GPU 封装进入正当时。在 AI 运算中，神经网络参数（权重、偏差、超参数和其他）需要存储在内存中，常规存储器与处理器之间的数据搬运速度慢，成为运算速度提升的瓶颈，且将数据搬运的功耗高。2016 年英伟达率先推出首款采用 CoWoS 封装的绘图芯片，为全球 AI 热潮拉开序幕。英伟达 H100 拥有 800 亿个晶体管，相比上一代的 A100，有着六倍的性能提升以及两倍的 MMA 改进，采用的 CoWoS 2.5D 晶圆级封装。在算力芯片性能暴增的时代下，相关的封装产业链也逐渐的进入高速发展时期。

Chiplet 是后摩尔时代的半导体工艺发展方向之一。Chiplet 将大型单片芯片划分为一组具有单独功能的小芯片单元 die（裸片），小芯片根据需要使用不同的工艺节点制造，再通过跨芯片互联和封装技术进行封装级别集成，降低成本的同时获得更高的集成度。

图表1：使用中介层的基于 Chiplet 的架构



资料来源：Semiconductor Engineering、方正证券研究所整理

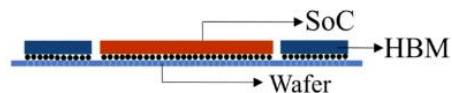
Chiplet 技术要把原本单个大硅片“切”成多个再通过封装重新组装起来，而单个硅片上的布线密度和信号传输质量远高于 Chiplet 之间，这就要求必须发展出高密度、大带宽布线的先进封装技术，尽可能的提升在多个 Chiplet 之间布线的数量并提升信号传输质量。支持 Chiplet 的底层封装技术目前主要由台积电、日月光、英特尔等公司主导，包含从 2D MCM 到 2.5D CoWoS、EMIB 和 3D Hybrid Bonding。

1.1 CoWoS：适用于 HPC 与 AI 计算领域的 2.5D 封装技术

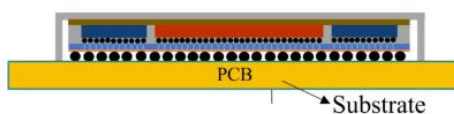
CoWoS（Chip-on-Wafer-on-Substrate）是台积电主导的，基于

interposer（中介介质层）实现的 2.5D 封装技术。CoWoS 先将芯片通过 CoW 封装至 Wafer（硅晶圆），并使用硅载片上的高密度走线进行互联，再把 CoW 芯片与 Substrate（基板）连接，整合成 CoWoS，达到封装体积小、功耗低、引脚少的效果。

图表2： CoWoS 封装示意图



(CoW) Chip on Wafer示意图



CoW芯片与基板（Substrate）连接示意图

资料来源：奇普乐芯片、方正证券研究所整理

TSV（Through Silicon Via，硅通孔）是 CoMoS 封装的关键技术。TSV 在芯片和芯片之间、晶圆和晶圆之间制作垂直导通，通过铜、钨、多晶硅等导电物质的填充，实现硅通孔的垂直电气互连，是目前唯一的垂直电互联技术。台积电根据中介层的不同，将其 CoWoS 封装技术分为三种类型：CoWoS-S、CoWoS-R、CoWoS-L。

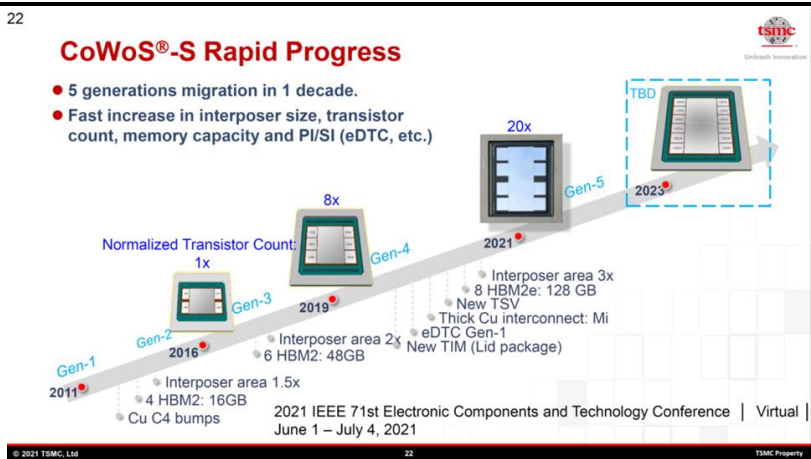
图表3： 台积电拥有的 CoWoS 封装技术

类型	中介层	主要方案	图示
CoWoS-S	Silicon 硅衬底	标准方案，基于硅中介层集成SoC和HBM的2.5D先进封装。	
CoWoS-R	RDL 重新布线层	引入了InFO技术中的RDL，RDL 中介层由聚合物和铜迹线组成，具有相对机械柔韧性，而这种灵活性增强了封装连接的可靠性，并允许新封装可以扩大其尺寸以满足更复杂的功能需求，从而有效支持多个Chiplets之间进行高速可靠互联。	
CoWoS-L	LSI & RDL 本地硅互联和重新布线层	<ul style="list-style-type: none"> 在每个产品中可以具有多种连接架构（例如 SoC 到 SoC、SoC 到小芯片、SoC 到 HBM 等），也可以重复用于多个产品，提供更灵活和可复用的多芯片互联架构。 中介层正面和背面均具有宽间距的 RDL 层，以及用于信号和功率传输的 TIV（中介层通孔）可在高速传输中提供低损耗的高频信号。 在 SoC 裸片下方集成其他元件，例如独立 IPD（集成无源器件），以支持其具有更好 PI/SI 的信号通信。 	

资料来源：台积电官网、WCCFTech、方正证券研究所整理

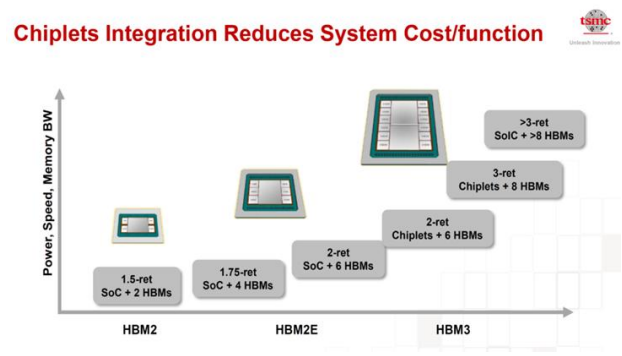
CoWoS-S 从 2011 年的第一代升级到 2021 年的第五代，第六代技术有望于 2023 年推出，将会在基板上封装 2 颗运算核心，同时可以板载多达 12 颗 HBM 缓存芯片。第五代 CoWoS-S 技术使用了全新的 TSV 解决方案，更厚的铜连接线，晶体管数量是第 3 代的 20 倍。它的硅中介层扩大到 2500mm^2 ，相当于 3 倍光罩面积，拥有 8 个 HBM2E 堆栈的空间，容量高达 128 GB。并且，台积电以 Metal TIM 形式提供最新高性能处理器散热解决方案，与第一代 Gel TIM 相比，封装热阻降低至 0.15 倍。

图表4： 台积电 CoWoS 封装的技术发展路线图



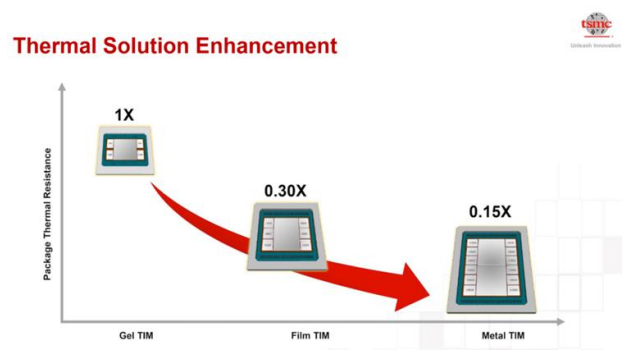
资料来源：台积电、WCCFTech、方正证券研究所整理

图表5： 第五代 CoWoS-S 可板载 8 颗 HBM 芯片



资料来源：台积电、WCCFTech、方正证券研究所整理

图表6： 第五代 CoWoS-S 热阻降低至 0.15 倍




资料来源：台积电、WCCFTech、方正证券研究所整理

AI 时代下算力需求日益增长，GPU 先进封装的重要性凸显。CoWoS 协助台积电拿下英伟达、AMD、Google 等高性能计算芯片订单。根据 DIGITIMES 报道，ChatGPT 日益普及所刺激的高端 AI 芯片需求激增，预计将推动对台积电 CoWoS 封装的需求，微软已与台积电及其生态系统合作伙伴接洽，商讨将 CoWoS 封装用于其自己的 AI 芯片。

英伟达高端 GPU 都采用 CoWoS 封装技术，将 GPU 芯片和 HBM2 集合在一起。2016 年英伟达推出 Tesla P100，通过加入采用 HBM2 的 CoWoS 第三代技术，将计算性能和数据紧密集成在同一个程序包内，提供的内存性能是 NVIDIA Maxwell 架构的三倍以上。并且，面向 HPC 和 AI 训练，英伟达以 Volta、Ampere 架构为基础推出了 V100、A100 高端 GPU，均采用台积电 CoWoS 封装，制程分别为 12nm、7nm，分

别配备 32 GB HBM2、40GB HBM2E 内存。基于台积电最先进的 CoWoS 封装,全新 Hopper 架构的 H100 GPU 制程达到 4nm,具有 80GB 的 HBM3 内存和超高的 3.2TB/s 内存带宽。

图表7: 英伟达高端 GPU 都采用 CoWoS 封装

类型	NVIDIA Tesla P100	NVIDIA Tesla V100	NVIDIA A100	NVIDIA H100
GPU	GP100	GV100	GA100	GH100
制程	16nm	12nm	7nm	4nm
晶体管数量	1534亿	2114亿	5404亿	8004亿
内存	16GB	32GB	40GB	80GB
内存类型	HBM2	HBM2	HBM2E	HBM3
带宽	720 GB/s	897.0 GB/s	1555GB/s	3.2TB/s
封装方式	CoWoS			
图示				

资料来源: 英伟达官网、方正证券研究所整理

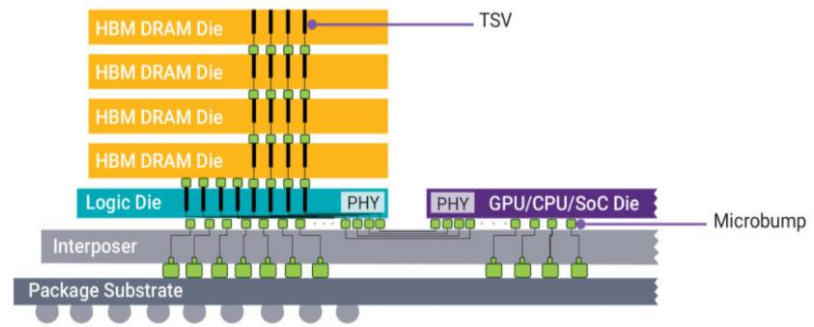
AMD 的数据中心加速器芯片将重新采用 CoWoS 封装。AMD 在 2017 年考虑将 Vega 20 的供应商从 GlobalFoundries 更换为台积电,主要看重其 7nm 工艺和 CoWoS 先进封装, Vega 20 配备 32GB HBM2 内存,直接对标英伟达 V100 加速器。根据 DIGITIMES 报道, AMD MI 200 原本由日月光集团与旗下矽品提供,应用 FO-EB 先进封装(扇出嵌入式桥接),新 MI 系列数据中心加速器芯片将重新采用台积电先进封装 CoWoS。基于 Aldebaran GPU 的 MI250 或采用第五代 CoWoS 封装技术,制程 6nm,实现 128GB HBM2E 内存等超高性能配置。

1.2 HBM: 存算一体化下的主流,突破了内存容量与带宽瓶颈

HBM 是“GPU+存储器”的模式,将解决高算力 AI 背景下芯片的“存算一体”问题。HBM (High Bandwidth Memory, 高带宽内存)是一款新型的 CPU/GPU 内存芯片,将多个 DDR 芯片堆叠在一起后和 GPU 封装在一起,实现大容量,高位宽的 DDR 组合阵列。HBM 主要是通过 TSV 技术进行芯片堆叠,即 DRAM 芯片上搭上数千个细微孔并通过垂直贯通的电极连接上下芯片; DRAM 下面是 DRAM 逻辑控制单元,对 DRAM 进行控制;GPU 和 DRAM 通过 uBump 和 Interposer(起互联功能的硅片)连通;Interposer 再通过 Bump 和 Substrate(封装基板)连通到 BALL;最后 BGA BALL 连接到 PCB 上。

虽然多核(例如 CPU)/众核(例如 GPU)并行加速技术也能提升算力,但在后摩尔时代,存储带宽制约了计算系统的有效带宽,芯片算力增长步履维艰,因此存算一体的芯片应运而生。存算一体是在存储器中嵌入计算能力,以新的运算架构进行二维和三维矩阵乘法/加法运算。存算一体的优势是打破存储墙,消除不必要的数据搬移延迟和功耗,并使用存储单元提升算力,成百上千倍的提高计算效率,降低成本。

图表8: HBM 堆叠结构



资料来源: Synopsys、方正证券研究所整理

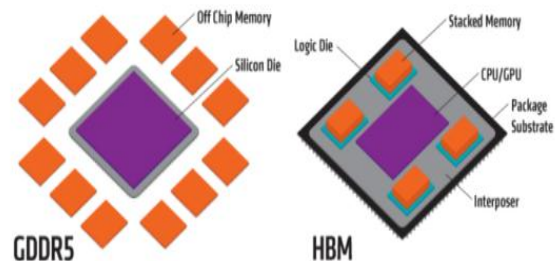
HBM 突破了内存容量与带宽瓶颈。凭借 TSV 方式, HBM 使 DRAM 从传统 2D 转变为立体 3D, 比 GDDR5 节省了 94% 的表面积, 随着半导体行业向小型化发展, HBM 能更充分地利用空间, 实现集成化。同时, HBM 大幅提高了容量和数据传输速率, 具有更高带宽、更多 I/O 数量、更低功耗, 革命性地提升了 DRAM 的性能。与 GDDR5 相比, GDDR5 内存每通道位宽 32bit, 带宽为 32GB/s; HBM2 的每个堆栈支持最多 1024 个数据 pin, 每 pin 的传输速率可以达到 2000Mbit/s, 那么总带宽为 256GB/s; 在 2400Mbit/s 的每 pin 传输速率之下, 一个 HBM2 堆栈封装的带宽就是 307GB/s。HBM 通过提升带宽、扩展内存容量, 提高了存储与 CPU/GPU 之间的数据传输速度, 从而减少了内存容量小带来的延迟问题。

图表9: 各封装技术参数对比

	DDR4	DDR5	HBM2	GDDR5	LPDDR4	LPDDR5
Applications	Servers → PCs → consumer	Servers → PCs → consumer	Graphics, HPC	Graphics	Mobile, auto, consumer	Mobile, auto, consumer
Typical interface (primary)	Server: 64+8 bits	Server: dual channel, 32+8 bits	Octal channel, 128-bit (1024 bits total)	Multi-channel, 32-bits	Mobile: quad channel, 16-bit (64-bits total)	Mobile: quad channel, 16-bit (64-bits total)
Typical interface (secondary)	Consumer: 32 bits	Consumer: 32 bits	None	None	Dual channel, 16-bit (32-bits total)	Dual channel, 16-bit (32-bits total)
Max Pin BW	3.2 Gb/s	6.4 Gb/s	2.0 → 2.4 Gb/s	8Gb/s	4.267Gb/s	6.4Gb/s
Max I/F BW	25.6 GB/s	51 GB/s	307 GB/s	32 GB/s	34 GB/s	51 GB/s
# Pins/channel	~380 pins	~380 pins	~2,860 pins	~170 pins	~350 pins	~370 pins
Max capacity	30S RDIMM: 128GB	30S RDIMM: 256GB	4H Stack: 4GB	One channel: 1GB	4 channels: 2GB	4 channels: 4GB
Peak volumes	*****	*****	**	*	*****	*****
Price per GB	\$	\$\$	\$\$\$\$	\$\$\$	\$\$	\$\$

资料来源: Synopsys、方正证券研究所整理

图表10: GDDR5 与 HBM 的外观对比



资料来源: 闪存市场、方正证券研究所整理

HBM3 即将问世, 最高的数据传输速率提升到 8.4Gbps。从 HBM 性能的历史演进来看, 2013 年, SK 海力士在业界首次成功研发出 HBM, HBM1 的数据传输速率大概可以达到 1Gbps 左右; 2016 年推出的 HBM2 为每个堆栈包含最多 8 个内存芯片, 同时管脚传输速率翻倍达 2Gbps; 2018 年推出的 HBM2E, 最高数据传输速率可以达到 3.6Gbps, 可实现每堆栈 461GB/s 的内存带宽。2021 年, SK 海力士和 Rambus 先后发布最高数据传输速率 6.4Gbps 和 8.4Gbps 的 HBM3 产品, 每个堆栈将提供超过 819GB/s 和 1075GB/s 的传输速率。SK 海力士 HBM3 显存的样品已通过 NVIDIA 的性能评估工作, 在 2022 年 6 月向 NVIDIA 正式供货; Rambus HBM3 或将在 2023 年流片, 实际应用于数据中心、AI、HPC 等领域。随着 HBM3 的性能提升, 未来市场空间广阔。

图表11: 各封装技术参数对比



资料来源：海力士、Rambus、方正证券研究所整理

2 相关标的

相关标的：长电科技、通富微电、华天科技、甬矽电子、晶方科技。

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格，保证报告所采用的数据和信息均来自公开合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响。研究报告对所涉及的证券或发行人的评价是分析师本人通过财务分析预测、数量化方法、或行业比较分析所得出的结论，但使用以上信息和分析方法存在局限性。特此声明。

免责声明

本研究报告由方正证券制作及在中国（香港和澳门特别行政区、台湾省除外）发布。根据《证券期货投资者适当性管理办法》，本报告内容仅供我公司适当性评级为C3及以上等级的投资者使用，本公司不会因接收人收到本报告而视其为本公司的当然客户。若您并非前述等级的投资者，为保证服务质量、控制风险，请勿订阅本报告中的信息，本资料难以设置访问权限，若给您造成不便，敬请谅解。

在任何情况下，本报告的内容不构成对任何人的投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需求，方正证券不对任何人因使用本报告所载任何内容所引致的任何损失负任何责任，投资者需自行承担风险。

本报告版权仅为方正证券所有，本公司对本报告保留一切法律权利。未经本公司事先书面授权，任何机构或个人不得以任何形式复制、转发或公开传播本报告的全部或部分内容，不得将报告内容作为诉讼、仲裁、传媒所引用之证明或依据，不得用于营利或用于未经允许的其它用途。如需引用、刊发或转载本报告，需注明出处且不得进行任何有悖原意的引用、删节和修改。

公司投资评级的说明：

强烈推荐：分析师预测未来半年公司股价有20%以上的涨幅；

推荐：分析师预测未来半年公司股价有10%以上的涨幅；

中性：分析师预测未来半年公司股价在-10%和10%之间波动；

减持：分析师预测未来半年公司股价有10%以上的跌幅。

行业投资评级的说明：

推荐：分析师预测未来半年行业表现强于沪深300指数；

中性：分析师预测未来半年行业表现与沪深300指数持平；

减持：分析师预测未来半年行业表现弱于沪深300指数。

地址	网址： https://www.foundersc.com	E-mail:yjzx@foundersc.com
北京	西城区展览馆路48号新联写字楼6层	
上海	静安区延平路71号延平大厦2楼	
深圳	福田区竹子林紫竹七道光达银行大厦31层	
广州	天河区兴盛路12号楼 隽峰苑2期3层方正证券	
长沙	天心区湘江中路二段36号华远国际中心37层	