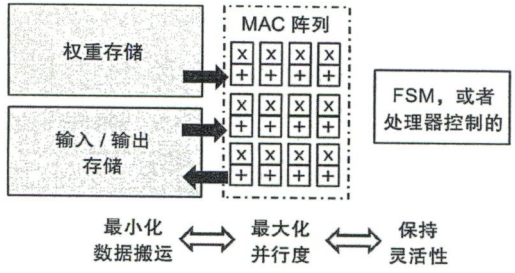

图 3-2　定制化的深度神经网络处理器通过最小化数据搬运和最大化并行度来提高效率。同时，在映射各种网络的同时保持灵活性是设计的关键所在。FSM 代表有限状态机

如果不利用数据在时间和空间中的局部性，试图向所有这些功能单元并发提供数据几乎是不可能实现的。实际上，一个网络层中的许多计算共享公共的输入。更具体地说，如图 1-5 中“朴素”CNN 的伪代码的醒目文字所示，每个权重参数在输出张量中同一切片的多个卷积计算中可以被重用大概 M^2 次，并且每个输入数据点都会被 F 个不同的输出张量的切片重用。此外，中间累加结果 o 需要累加 $C\times k^2$ 次。定制化加速器可以通过多种方式来利用这些数据重用性以进一步提高效率，而这是具备高度并行但是没有数据流优化的 GPU 所做不到的。

一方面，数据复用可以通过在多个并发的执行单元上复用同一数据，或等效地在同一个执行单元的不同时隙中复用数据来实现。在这种拓扑中，可以区分为 3 种极端情况：

- 在“权重并行”或“**输入固定**”的方法中（见图 3-3），同一输入数据会跟同一层中不同输出通道的若干权重相乘。在理想情况下，这里每个输入将只加载一次到系统中。但是这会对权重的存储带宽（BandWidth，BW）产生负面影响，因为每次产生新的输入都需要频繁地重新加载权重。而且输出结果的累加无法在不同的时钟周期完成，需要将中间结果缓存在存储器中以便之后重新取回，这严重影响了输入 / 输出（I/O）存储器的带宽。
- “**权重固定**”或“输入并行”的方法改善了权重存储带宽，但以输入存储带宽为代价。这里每个权重都被提取一次并与许多输入值相乘。