1. 冯·诺依曼型计算机的主要设计思想是什么？它包括哪些主要组成部分？

答：冯·诺依曼计算机的基本思想是：存储程序控制，将用指令序列描述的解题程序与原始数据一起，存储到计算机中。计算机只要一启动，就能自动地依次取出一条条指令并执行，直至程序执行完毕，得到计算结果为止。

按此思想设计的计算机硬件系统包含：运算器、控制器、存储器、输入设备和输出设备 5 个基本部件。运算部件的作用是用来进行数据变换和各种运算；控制部件则在计算机提供统一的时钟下，把程序中的各基本操作进行时序分配，并发出相应的控制信号，驱动计算机各部件按节拍有序地完成程序规定的操作内容；存储器用来存放程序、数据及运算结果；输入输出设备的主要作用是接受用户提供的外部信息或向用户提供输出信息。

2. 指令和数据均存放在内存中，CPU 如何从时间和空间上区分它们是指令还是数据？

答：时间上讲，取指令事件发生在取指周期，取数据事件发生在执行周期。

空间上讲，从内存读出的指令流向控制器；从内存中读出的数据流一定流向运算器。

3. 说明计算机系统的层次结构。

答：计算机系统层次结构为：微程序机器级、一般机器级（或称机器语言级）、操作系统级、汇编语言级、高级语言级。

4. 试述浮点数规格化的目的。

答：浮点的规格化是为了使浮点数尾数的最高数值位为有效数位。当尾数用补码表示时，若符号位与小数点后的第一位不相等，则被定义为已规格化的数，否则便是非规格化数。通过规格化，可以保证运算数据的精度。

5. 运算器由哪几部分组成？

答：运算器的基本结构应包括以下几个部分：

（1）能实现算术和逻辑运算功能的部件 ALU；

（2）存放待加工的信息或加工后的结果信息的通用寄存器组；

（3）按操作要求控制数据输入的部件：多路开关或数据锁存器；

（4）按操作要求控制数据输出的部件：输出移位和多路开关；

（5）运算部件与其它部件进行信息传送的总线以及总线接收器与发送器。

6. 计算机系统中，存储器为什么要采用多级存储器体系结构？简述理由。

答：计算机存储器子系统一般分为高速缓存、内存和外存。高速缓存 Cache 是为了解决 CPU 与主存之间的速度不匹配问题而设置的，其性能是速度快、容量小；内部存储器（即主存）容量大、速度较慢(相对于 Cache)，通常用于存放运行的程序和数据；外部存储器容量巨大，可读可写，单位存储成本最低，且可以脱机保存信息。现代微机把这些不同容量、不同速度的存储器按一定的体系结构组织起来，形成一个统一的存储系统，主要是为了解决存储容量、存取速度和价格之间的矛盾。

7. 动态 MOS RAM 为何要刷新？有哪几种刷新方式？各有何特点。

答：动态存储元是依靠栅极电容上有无电荷来表示信息的，但电容的绝缘电阻不是无穷大，因而电荷会泄露掉。为了使已写入存储器的信息保持不变，一般每隔一定时间必须对存储体中的所有记忆单元的栅极电容补充电荷，这个过程就是刷新。

（1）集中刷新：前一段时间进行读/写/保持，后一段进行集中刷新，缺点是在集中刷新的这段时间内不能进行存取访问，称之为死时间。

（2）分散或异步刷新：首先用刷新的行数对刷新周期（如 2ms）进行分割，然后将已分割的每段时间分为两部分，前段时间用于读/写/保持，后一小段时间用于刷新，这样既充分利用了 2ms 的时间，又能保持系统的高速性。

8. Cache 与主存之间的地址映射方法有哪几种？各有何特点？

答：Cache 与主存之间的地址映射方法有全相联映射、直接映射、组相联映射三种。全相联映射是每个主存块可以映射到任何一个 Cache 块中，最灵活但实现的成本代价最大；直接映射是每个主存块只能映射到一个唯一对应的 Cache 块中，实现简单但 Cache 利用率低；组相联映射是每个主存块唯一对应一个 Cache 组，但可映射到组内任何一个块中，是前两种方式的折中。

9. 在什么情况下，会发生 Cache 和主存对应单元的内容不一样？一般采取哪些方法解决 Cache 和主存的不一致问题？

答：如果程序执行过程中要对该字块的某个单元进行写操作，就会遇到 Cache 与主存的不一致性问题。通常有两种方法解决：一种方法是暂时只向 Cache 存储器写入，并用标志加以注明，直到经过修改的字块被从 Cache 中替换出来是才一次写入主存；第二种方式是每次写入 Cache 存储器时也同时写入主存，使 Cache 和主存保持一致。当被修改的单元不在 Cache 存储器时，写操作直接对主存进行，而不写入 Cache 存储器。

10. Cache 和虚拟存储器在原理和功能方面有何相同和不同之处？

答：相同之处：都利用了程序局部性原理，把程序划分为许多信息块，运行时自动地把信息块从慢速存储器向快速存储器调度，信息块的调度都采用一定的替换策略以提高继续运行时的命中率。另外都是为了提高性能价格比。

不同之处：Cache 用于弥补主存与 CPU 之间的速度差异，而虚拟存储器则用于弥补主存容量的不足；Cache 每次传送的信息块是定长的，且只有几十字节，虚拟存储器的信息块可以是定长的（页），也可以是不定长的（段），长度也比较大；CPU 可直接访问 Cache，但不能直接访问辅存；Cache 与主存信息交换的过程全部由硬件实现，主存与辅存的信息交换则通过辅助硬件与存储管理软件来完成。

11. 试比较内存、外存、缓存、控存、虚存。

答：

（1）内存：直接和 CPU 交换信息。

（2）外存：是为了扩大内存容量的辅助存储器，不直接和 CPU 交换信息，容量比内存大，速度比内存慢。

（3）缓存：是为了解决内存和 CPU 的速度匹配、提高访存速度的一种存储器，它设在内存和 CPU 之间，速度比内存快，容量比内存小，存放 CPU 最近期要用的信息。

（4）控存：是微程序控制器中用来存放微指令的存储器，由 ROM 组成，速度比主存更快。

（5）虚存：是为了解决扩大主存容量和地址分配问题提出的存储器模型，把主存和辅存统一成一个整体。从整体上看，速度取决于主存，容量取决于辅存，实际上 CPU 仍然只与主存交换信息，由操作系统和硬件共同实现主存和辅存之间信息的自动交换。

12. 一个较完善的指令系统应包括哪几类？

答：包括数据传送指令、算术运算指令、逻辑运算指令、程序控制指令、输入输出指令、堆栈指令、字符串指令、特权指令等。

13. 为什么要引入指令系统？有哪些常见的寻址方式？对于一个指令系统来说，如果使用复杂的寻址方式，对 CPU 的结构、性能方面各有什么影响？

答：引入指令系统后，首先，可以避免用户与二进制机器码直接接触，使得用户编写程序更为方便。其次，指令系统是表征一台计算机性能的重要因素，他的格式和功能不仅直接影响到机器的硬件结构，而且也直接影响到系统软件，影响到机器的适用范围。常见的寻址方式有：①隐含寻址、②立即寻址、③直接寻址、④间接寻址、⑤寄存器寻址、⑥寄存器间接寻址、⑦相对寻址、⑧基址寻址、⑨变址寻址。复杂的寻址方式会使 CPU 的结构同样的变复杂，不利于流水线的运行，降低 CPU 运行效率。

14. 简述 CPU 的主要功能。

答：CPU 主要有以下四方面的功能：

（1）指令控制：程序的顺序控制，称为指令控制。

（2）操作控制：CPU 管理并产生由内存取出的每条指令的操作信号，把各种操作信号送往相应部件，从而控制这些部件按指令的要求进行动作。

（3）时间控制：对各种操作实施时间上的控制，称为时间控制。

（4）数据加工：对数据进行算术运算和逻辑运算处理，完成数据的加工处理。

15. 举出 CPU 中 6 个主要寄存器并说明其名称及功能。

答：（1）指令寄存器（IR）：用来保存当前正在执行的一条指令。

（2）程序计数器（PC）：用来确定下一条指令的地址。

（3）地址寄存器（AR）：用来保存当前 CPU 所访问的内存单元的地址。

（4）缓冲寄存器（DR）：<1>作为 CPU 和内存、外部设备之间信息传送的中转站。

<2>补偿 CPU 和内存、外围设备之间在操作速度上的差别。

<3>在单累加器结构的运算器中，缓冲寄存器还可兼作为操作数寄存器。

（5）通用寄存器（AC）：当运算器的算术逻辑单元（ALU）执行全部算术和逻辑运算时，为 ALU 提供一个工作区。

（6）状态条件寄存器：保存由算术指令和逻辑指令运行或测试的结果建立的各种条件码内容。除此之外，还保存中断和系统工作状态等信息，以便使 CPU 和系统能及时了解机器运行状态和程序运行状态。

16. 什么是指令周期？什么是机器周期？什么是时钟周期？三者之间的关系如何？

答：指令周期是取出并执行一条指令所需的时间，包括取指令、分析指令和执行指令所需的全部时间。机器周期也称为 CPU 周期，通常等于取指时间（或访存时间）。时钟周期是时钟频率的倒数，也可称为节拍脉冲或 T 周期，是处理操作的最基本单位。一个指令周期由若干个机器周期组成，每个机器周期又由若干个时钟周期组成。

17. 简述计算机常见指令运行的三个阶段?

答：常见指令运行的三个阶段包括：①取指周期：根据 PC 中的内容按照一定寻址方式从主存中取出指令代码并放在 IR 中；②间指周期或分析周期：根据指令字中的地址码取操作数；③执行周期：根据 IR 中指令字的操作码和操作数通过 ALU 操作产生结果。

18. 微程序控制器的工作原理。

答：微程序控制器的基本原理是仿照通常的解题程序的方法，把操作控制信号编成所谓的“微指令”，存放到一个只读存储器里。当机器运行时，一条又一条地读出这些微指令，从而产生全机所需要的各种操作控制信号，使相应部件执行所规定的操作。

19. 说明程序与微程序、指令与微指令的异同。

答：程序和微程序都可以用程序设计的方法进行设计。其区别是前者由机器指令组成，存于存储器；而后者由微指令组成，存于控制存储器，一个微程序对应一条机器指令。

指令和微指令都是计算机的操作命令。但前者存于存储器，由操作码和地址码两部分组成。操作码经译码后与时序、状态条件等组合产生微操作，其地址码部分是用来给出操作数地址的。微指令控制存储器，包含一组微命令，经组合后可产生一组微操作，它还包含地址字段，但这个地址是用来确定下一条要执行的微指令的地址。

20. 控制器的设计有哪两种方法？这两种方法有什么异同点？

答：微程序控制和组合逻辑控制。

组合逻辑控制器与微程序控制器相同之处都是根据指令操作码和时序信号，产生各种控制信号，以便正确地建立各种数据通路，完成取指令和执行指令的控制。

组合逻辑控制器的优点是由于控制器的速度取决于电路延迟，所以速度较快；缺点是由于控制器部件看成专门产生固定时序控制信号的逻辑电路，所以把用最少元件和取得最高速度作为设计目标，一旦设计完成，不能通过其他的修改添加新功能。

微程序控制器的优点同组合逻辑控制器相比，具有规整性、灵活性、可维护性等一系列优点；缺点是由于微程序控制器采用了存储程序原理，所以每条微指令都要从控制存储器中取出，从而影响了速度。

21. 用定量分析法说明流水处理器比非流水处理器具有更高的吞吐率。

答：（1）在流水线处理器中，一个具有 k 级过程段的流水线处理 n 个任务时，需要的时钟周期数为：T2=k+（n-1），其中 k 个时钟周期处理第一个任务，k 个周期后，流水线被填满，剩余的（n-1）个任务只需（n-1）个时钟周期就可完成。

（2）当用非流水处理器来处理这 n 个任务时，因串行方式工作，所需的时钟周期数为：T1=n*k，由此得 k 级流水线处理器的加速比为：Ck=T1/T2=n*k/[k+（n-1）]，当 n>>k 时,Ck→k.这就是说,理论上 k 级流水线处理器几乎可以提高 k 倍的速度,因而比非流水处理器具有更高的吞吐率。

22. 简单列出影响 CPU 流水线的因素，并针对每一种因素的不同类型给出解决方法。

答：① 资源冲突（结构冒险）：1）前一指令访存时，使后一条相关指令（及其后续指令）暂停一个时钟周期；2）单独设置数据存储器和指令存储器。

②数据冲突（数据冒险）：1）把数据相关指令及其后续指令都暂停一至几个时钟周期，直到数据相关问题消失后再继续执行，可以分为硬件阻塞（stall）和软件插入 NOP 指令两种方法。2）设置相关专用通路，即不等前一条指令把计算结果写回寄存器，下一条指令也不再读寄存器，而直接把前一条指令的 ALU 计算结果作为自己的输入数据开始计算过程。即数据旁路技术。3）通过编译器对数据相关的指令编译优化的方法，调整指令执行顺序来解决数据相关。

③控制冲突（控制冒险）：1）对转移指令进行分支预测，尽早生成转移目标地址。2）预取转移成功和不成功两个控制流方向上的目标指令。3）加快和提前形成条件码。4）提高转移方向的猜准率。

23. 什么是总线？以总线组成计算机有哪几种组成结构？系统总线按其传送的信号可分为哪几类？

答：总线是构成计算机系统的互联机构，是多个系统功能部件之间进行数据传送的公共通路。按照总线的连接方式，计算机组成结构可以分为单总线结构、双总线结构和多总线结构等。系统总线按其传送的信号可分为地址线、数据线、控制线。

24. 简答在计算机总线中，为什么同步定时方式有较高的传输速率？ （与异步定时方式相比）

答：同步定时方式是指系统采用一个统一的时钟信号来协调发送和接收双方的传送定时关系。时钟产生相等的时间间隔，每个间隔构成一个总线周期。在一个总线周期中，发送方和接收方可以进行一次数据传送。因此采用统一的时钟，每个部件或设备发送或接收信息都在固定的总线传送周期中，总线控制逻辑简单，具有较高的传输速率。

25. 计算机 I/O 数据传送控制方式通常可分为哪几种？

答：程序查询方式、程序中断方式、直接内存访问（DMA）方式、通道方式、外围处理机方式。

26. 简要说明中断响应和中断处理的过程。

答：当 CPU 执行完一条现行指令时，如果外设向 CPU 发出中断请求，那么 CPU 在满足响应条件下，将发出中断响应信号，同时关中断（不再受理另外设备的中断），并寻找中断源，保护现场（包括断点、累加器、寄存器等），转向中断服务程序，结束后关中断，恢复现场，返回主程序，再开中断。

27. 中断响应优先级与中断处理优先级的区别？

答：中断响应优先级是由硬件排队线路或中断查询程序的查询顺序决定的，不可以动态调整。中断处理优先级是 CPU 实际响应中断请求的优先顺序，可通过中端屏蔽字去改变优先级。反映的是正在处理的中断是否比新发生的中断的处理优先级低（屏蔽位为‘0’，对新中断开放），如果是的话，就中止正在处理的中断，转到新中断去处理，处理完后再回到刚才被中止的中断继续处理。

28. 查询方式和中断方式的主要异同点是什么？

答：两种方式都是以 CPU 为中心的控制方式，都需要 CPU 执行程序来进行 I/O 数据传送。程序查询方式控制简单，但系统效率很低，无法实现并行操作；中断方式通过服务程序完成数据交换，实现了主机与外设的并行性。

29. 为什么在计算机系统中引入 DMA 方式来交换数据？若使用总线周期挪用方式，DMA 控制器占用总线进行数据交换期间，CPU 处于何种状态？

答：DMA-直接内存存取。引入的目的是为了减轻 CPU 对 I/O 操作的控制，使得 CPU 的效率有了提高。可能遇到两种情况：一种是此时 CPU 不需要访内，如 CPU 正在执行乘法命令；另一种情况是 I/O 设备访内优先，因为 I/O 访内有时间要求，前一个 I/O 数据必须在下一个访内请求到来之前存取完毕。

30. 在输入输出系统中，DMA 方式是否可以替代中断方式？试比较 DMA 方式和中断控制方式。

答：DMA 方式不可以替代中断方式，因为在 DMA 方式的传送结束阶段，还需要向 CPU 发出中断请求。DMA 方式与中断方式的主要区别有：

（1）中断方式通过程序实现数据传送，而 DMA 方式不使用程序，直接靠硬件来实现。

（2）CPU 对中断的响应是在执行完一条指令之后，而对 DMA 的响应则可以在指令执行过程中的任何两个存储周期之间。

（3）中断方式不仅具有数据传送能力，而且还能处理异常事件；DMA 只能进行数据传送。

（4）中断方式必须切换程序，要进行 CPU 现场的保护和恢复操作；DMA 仅挪用了一个存储周期，不改变 CPU 现场。

（5）DMA 请求的优先权比中断请求高。CPU 优先响应 DMA 请求，是为了避免 DMA 所连接的高速外设丢失数据。

31. 人工智能、大数据技术是行业内关注的新焦点，这些技术的发展对计算机 CPU 处理器和计算机架构提出了新的需求，例如深度学习算法需要海量数据的训练，而传统计算架构无法支撑深度学习算法的大规模计算需求。请从存储系统、指令系统、处理器架构三个方面分别给出一些设计优化的思路，以适应智能学习算法所具有的大计算量、高数据带宽、数据类型复杂等特点。

答：针对智能学习算法，存储系统可以设计为多级存储，内存计算等；指令系统可以针对该类算法设计专用指，设计特殊寻址方式等；处理器架构采用乱序执行，超流水线，SIMD, VLIW 等技术。