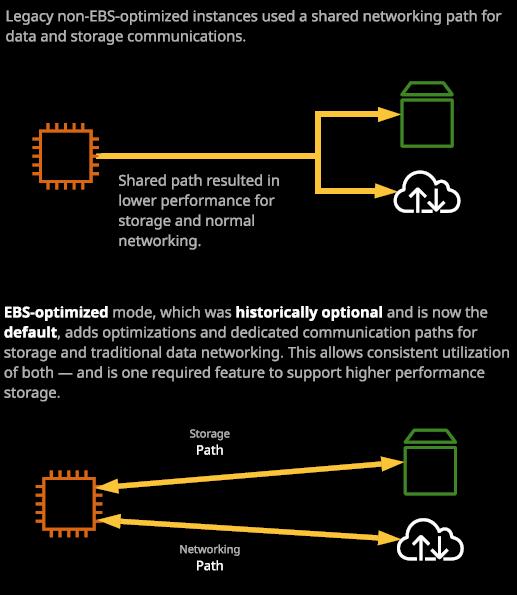
**EBS OPTIMIZATION, ENHANCED NETWORKING AND PLACEMENT GROUPS**



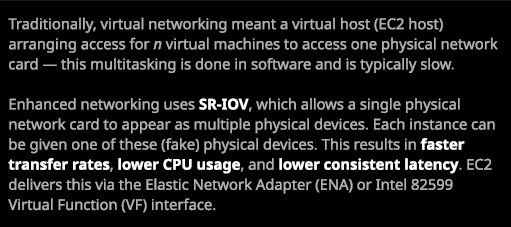
We'll recover three different topics that all go together to govern the performance of an EC2 instance. The first thing we're going to cover is EBS optimization. The second is enhance networking, and the final subject we'll cover in this lesson is placement groups.

Historically, the way the EC2 instances communicated with both storage and networking was over a shared communication path. So a single network interface on the EC2 host was used by one or more EC2 instances on that shared communication into phase carried both storage data as well as traditional networking data. Now the problem with that is that because it was a shared communication path, both of those different types of data contested for access to that storage path. So if you had an application that was ruling on an EC2 instance that was demanding, if it's storage, you'd find that could also disrupt traditional networking transfer and the same was true in reverse. If you had an application that was heavy on normal data transfer, it could impact storage performance **EBS optimization** was one of the initial performances optimization features added to EC2 and it created a separate, dedicated path for storage. So if you enabled EBS optimize mode, which was historically optional, then you get access to a second path for data transfer. So no longer would your normal networking path and your storage data transfer contend with each other that operate over individual paths and that provided three different benefits. **The first was access to better and faster storage.** **The second was improved network data transfer rates and the third was a higher level of consistency on both of those, so you wouldn't get a large amount of performance variants as both of these contenders for the same data transfer path.**

Now, historically, this was optional. It was an option available at this stage you could launch. It is an EBS optimized instance and originally it did cost extra. You needed to specifically add it and then and only then would you gain the performance benefits. Now, at this point, all modern instances include this by defaults. It comes as default and you don't get any performance change or pricechange by disabling it. **In order to get maximum performance from EBS you need to select the appropriate sized instance, because, of course, larger instances come with higher performance levels but you also need to make sure that using EBS optimized instances which any current generation instances because with those you get the maximum performance.**

Now, another performance technique you need to be aware of is to do with snapshots. When you initially launched an EC2 instance, or when you manually launch an EBS volume, you automatically get the maximum performance from that volumes. If I create ah 100 GB volume then straight away, I'm going to get the maximum performance from that volume that I should based on the storage type that I selected. Now, if I create a snapshot from this EBS volume and I'll just put some random text in for the name and hit create snapshot. At this point, the data on this volume has been copied into S3, and it forms part of the snapshot. So this is a snapshot that's just been completed. Now, if this snapshot contained 100 GB of data and to create a new volume from the snapshot, what it would actually do would be to create the new value and allocate the space EBS but perform the copy off the data from S3 to EBS overtime. So when you create the initial volume, but you base it on a snapshot, it doesn't immediately copy all that data to EBS. It copies it in the background. **This means that you don't get the maximum performance of an EBS volume created from a snapshot until all that data has been copied across in the background. If you ask for some data from the volume that's not yet been copied, it will immediately be copied.** Now, there are techniques that you can do to improve this performance and one of those is **perform a read of every part of that volume in advance before you move into production** but just be aware that there are performance implications by restoring volumes from snapshots. So that's EBS optimization and just a little piece of information about snapshots.

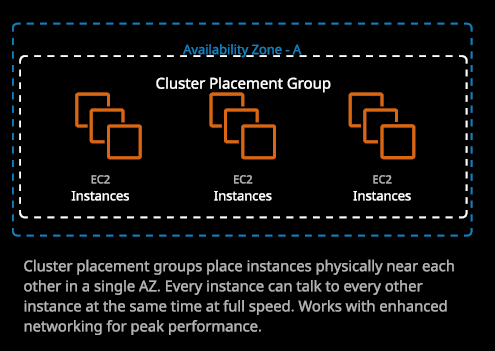
Now, the next two things I want to talk about enhance networking and placement groups. **EC2 is a virtualization product. It's essentially a virtual machine, known as an instance, that's running on an EC2 host.** An EC2 host is physical hardware and running on the EC2 host is a **piece of software called a hypervisor, and it's the responsibility of this hypervisor to carve up the physical resources, present them as individual virtual machines, and then control access to the physical resources for those virtual machines.** So the **hypervisor is the thing that controls, whether it's your EC2 instance or my EC2 instance, that currently has access to the networking card that's in the physical EC2 host.** Now historically, that was done with software. It was efficient software, and it's optimized but it's still emulating real hardware, so your virtual machine, your EC2 instance it's not accessing its own dedicated networking card. It's running through software, and that comes with some performance implications because its software it's not as fast as accessing through real hardware, and it's not as consistent.



Now **enhanced networking, which is a feature of EC2 instances, uses a technique called SRIOV and this stands for single root input output virtualization.** **One physical network adapter in the EC2 host is capable as presenting itself as multiple kind of fake physical adapters, and these are known as virtual functions. So rather than the one physical network card presenting itself as one physical adapter, it presents itself as multiple physical adapters. For a hypervisor that supports this, it means that each virtual machine can be given direct access to one of these virtual functions. So rather than the hypervisor having to control which virtual machine accesses hardware at what time the virtual machines or EC2 instances are capable of interacting directly and it's the network card, which handles this interaction.** Now this comes with a few benefits. **It means less CPU is used on the host during heavy utilization. It means lower latency levels, and it means a better consistency of latency so it doesn't vary, very much. Essentially, it gives us better networking performance. So enhanced networking when you're using EC2 uses SRIOV. It's what allows the higher networking performance better latency and all of the good networking things.**

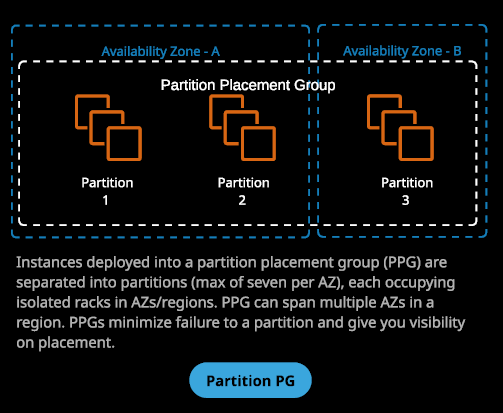
Now EC2 provides **two network adapters** in instances that provide this advanced functionality, the first is the **Intel virtual adapter** and the second is the **Elastic Network Adapter**, or ENA which provides really high end speeds and a lot of the newer EC2 instance types and generations will be coming with this by default. **So enhance networking is one component of achieving maximum performance of networking inside AWS.**

The other **is placement groups now there are three types of placement groups available in EC2 and they all exist for different reasons**. At a high level, though, **they allow some level of control or visibility about where your EC2 instances run from, it essentially allows you to see or to control physical location.** Now that's powerful, because we can use **that to improve performance or reliability**.

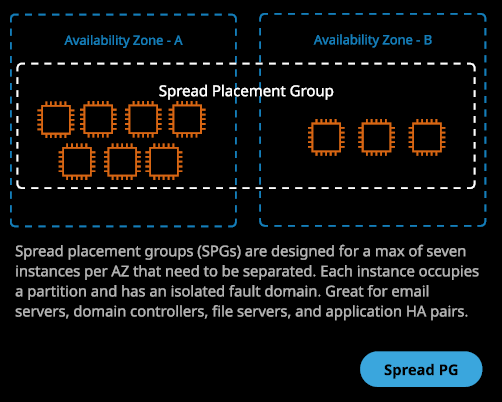


The first type of placement group is known as a **cluster placement group**. **Cluster placement groups are designed for performance, they're pretty easy to configure.** We just go to placement groups, create a placement group. We'll give it a name. So I'll call this one clusterpg for cluster placement group, and you'll be asked to select a strategy. Of course, there are **three strategies cluster, spread, and partition**. We're covering cluster so I'll select that and hit create and that's it. You don't need to do anything else when you create a placement group. Now, **cluster placement groups are purely designed for performance. They're limited to one availability zone**. So when you create a cluster placement group, you don't specify the availability zone but when you launch instances into that placement group, they're locked into the availability zone at the first instance that's launched into them but what they actually do **is they ensure that the instances launched inside a cluster placement group are physically close together.** So let's say you launched four EC2 instances into this placement group. What you'll tend to find is that these EC2 instances will probably be running on the same EC2 host so they'll achieve maximum network performance because that networking traffic might not even need to leave that host. It can occur internally, so that's always the best option and **cluster placement groups are the thing that will allow you to achieve the maximum possible performance inside** **EC2**. Any EC2 instances that are inside a placement group will be able to achieve the maximum possible networking speed between those EC2 instances, and they can all do so at the same time. **Cluster placement groups work best with enhance networking in order to get that peak performance.**

So keep that in mind for the exam, generally, **for any high performance instances, you, of course, need to pick a latest generation, which will include EBS optimization. You need to make sure you're using enhance networking, and you need to make sure that they're placed inside a placement group and if you do those you're going to get the maximum possible performance.** Now for the exam and real world usage a tip is that you should **always try to launch all of the instances that go inside a placement group at the same time because when you do that AWS reserve capacity. Now they do reserve extra capacity.** So, for example, if I launched four EC2 instances into this placement group, this clusterpg then AWS they're going to find a suitable occasion to run those four EC2 instances from. They might make sure this space for an additional two in case I want to launch more but it is possible if I only launch four originally and I try to launch another four, I might get a capacity issue. So there can be capacity issues that is solely isolated to a cluster placement group. **In general, launch all the instances that you want to launch initially with the placement group and additionally, best practice is to always pick the same type of instances to be in the same cluster placement group**. So, by picking the same type and launching them all initially, you get the maximum chance that you'll be able to launch the placement group with the instances successfully. So **cluster placement groups can only be in one availability zone, so that's the way that they achieve the maximum performance is to be in one availability zone.**



Next, we've got **partition placement groups. Now partition placement groups are designed to ensure maximum application availability**. Let's say you've got a particular application, maybe an analytics application or a big data application by using partition placement groups, you're able to ensure that instances are launched into separate partitions. So what's a partition? Well, **it's an isolated group of infrastructure inside AWS. So inside an availability zone, you might have multiple partitions**. Think of these as racks, so storage, compute, networking altogether and they're kind of isolated from the other partitions. The idea is, **if a single partition fails, you don't lose resources in another partition, so partition placement groups are a way that you get exposed to an even smaller fault domain than availability zones.** You might have an application that uses hundreds of EC2 instances, and you want to make sure that they're evenly split across all of the infrastructure partitions inside a region and so you'll create a partition placement group that can spread across different availability zones and each of those availability zones will have **seven partitions** in, and you can use that to split those instances across all of those seven partitions. So you get visibility of that and, if needed, you can give that information to applications so the application itself can have visibility over its infrastructure placement. So generally, **partition placement groups are designed to ensure availability and some level of infrastructure spread for applications that use a large amount of EC2 instances. So partition placement groups tend to only be use for larger infrastructure deployments** but it's worthwhile knowing exactly how they work.



**Spread placement groups are purely for availability. So spread placement groups are designed for a maximum of seven instances per availability zone.** There are seven partitions per availability zone. **With spread placement groups, it means each individual instance is going to be in its own partition. They're designed for small infrastructure deployment, so there's going to be a maximum of seven Instances per availability zone and they're design for situations when you want to ensure that every single instance is operating on a separate partition. So these great for email servers, domain controllers, file servers, and application, highly available pairs anything where you need to ensure that the failure of one partition won't take down all of your instances then you would use a spread placement group and that's the theory of placement groups.**

So you're going to use **cluster placement groups for maximum performance**, **spread placement groups for maximum availability, and then partition placement groups if you've got large infrastructure platforms where you want to have some visibility of where those instances are from a partition perspective.**