**Tempo de Resposta:**  
É o tempo total entre uma interação do usuário (tecla enter, tecla de função, cliques etc) e a próxima tela.

Achei interessante um exemplo de tempo de resposta de pixel de TV:

O tempo de resposta de um monitor ou televisão indica o intervalo necessário para um pixel ativo (preto) ficar inativo (branco) e voltar a ser ativado novamente (preto).

https://www.tecmundo.com.br/monitor/95072-tempo-resposta-atualizacao-monitor-afetam-games.htm

**Responsividade:**

É o tempo que um processamento aguarda para iniciar no servidor.

É o tempo de processamento de uma etapa de um programa que está sendo solicitado. Depende da lógica e do volume de coisas que estão sendo processadas;

Tem um exemplo no meu trabalho, podemos efetuar uma programação de horário para processamento de arquivos, mas nem sempre o servidor inicio no horário programado, pois há o tempo de receber a requisição.

https://pruvo.wordpress.com/2015/08/11/entenda-a-arquitetura-para-entender-o-tempo-de-resposta-do-seu-sap/

**Latência**:

Como exemplo podemos pegar a latência da nossa rede:

Em uma rede, latência é sinônimo de atraso, é uma expressão de quanto tempo leva para um pacote de dados  ir de um ponto designado para o outro (Do seu computador até este site por exêmplo). A latência é medida através do envio de um pacote que é devolvido para o remetente e o tempo de ida e volta é considerada a latência (ping).

https://brasilwork.com.br/duvidas/e-latencia-em-redes/

**Throughpu**

Em [redes de comunicação](https://pt.wikipedia.org/wiki/Redes_de_comunica%C3%A7%C3%A3o), como [Ethernet](https://pt.wikipedia.org/wiki/Ethernet) ou [packet radio](https://pt.wikipedia.org/w/index.php?title=Packet_radio&action=edit&redlink=1" \o "Packet radio (página não existe)), **throughput**, **throughput de rede** ou simplesmente [taxa de transferência](https://pt.wikipedia.org/wiki/Taxa_de_transfer%C3%AAncia) é a quantidade de dados transferidos de um lugar a outro, ou a quantidade de dados processados em um determinado espaço de tempo.

https://pt.wikipedia.org/wiki/Throughput

Carga:

A carga esta sempre atrelada a escabilidade:

A escalabilidade é um assunto extremamente importante em sistemas eletrônicos, [bancos de dados](https://pt.wikipedia.org/wiki/Banco_de_dados), [roteadores](https://pt.wikipedia.org/wiki/Roteador), redes de computadores, etc, e implica desempenho. Um sistema cujo desempenho aumenta com o acréscimo de [hardware](https://pt.wikipedia.org/wiki/Hardware), proporcionalmente à capacidade acrescida, é chamado "sistema escalável".

https://pt.wikipedia.org/wiki/Escalabilidade

**Sensibilidade a Carga:**

É o tempo de resposta que varia com a carga:

Digamos que o sistema A tenha um tempo de resposta de 0,5 segundo para u m número de usuários entre 10 e 20, e o sistema B tenha um tempo de resposta de 0,2 segundo para 10 usuários que aumenta para 2 segundos com 20 usuários.

Nesse caso o sistema A tem uma sensibilidade de carga menor que o sistema B.

https://books.google.com.br/books?id=8uSN0Qwz2kQC&pg=PA29&lpg=PA29&dq=tempo+de+resposta+conforme+a+carga&source=bl&ots=gqfNSuxaeH&sig=gOitLeXBWJwRc3Ke26tXxM26SXM&hl=pt-BR&sa=X&ved=0ahUKEwiWzJH6y\_HYAhUChuAKHUI0CAcQ6AEINzAD#v=onepage&q=tempo%20de%20resposta%20conforme%20a%20carga&f=false