K-Means Clustering for KDD Dataset

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```

In [2]:
```
# Load data
df_kddcup = pd.read_csv('kddcup_small.csv')
df_kddcup = df_kddcup.iloc[:, [0, 7, 10, 11, 13, 35, 37, 39]]
```

In [3]:
```
# Normalization
df_kddcup = (df_kddcup - df_kddcup.mean()) / df_kddcup.std()
```

In [4]:
```
kddcup_array = np.array([df_kddcup['duration'].tolist(),
                         df_kddcup['wrong_fragment'].tolist(),
                         df_kddcup['num_failed_logins'].tolist(),
                         df_kddcup['logged_in'].tolist(),
                         df_kddcup['root_shell'].tolist(),
                         df_kddcup['dst_host_same_src_port_rate'].tolist(),
                         df_kddcup['dst_host_serror_rate'].tolist(),
                         df_kddcup['dst_host_rerror_rate'].tolist(),
                         ], np.float)
kddcup_array = kddcup_array.T
```

In [5]:
```
# Clustering
CLUSTER_NUM = 5

model = KMeans(n_clusters=CLUSTER_NUM)
pred = model.fit_predict(kddcup_array)

df_kddcup['cluster_id'] = pred
```

In [6]:
```
print(df_kddcup['cluster_id'].value_counts())
```

```
1    40
4    30
0    30
2    29
3    21
Name: cluster_id, dtype: int64
```

In [17]:
```
# Visualization using Matplotlib

cluster_info = pd.DataFrame()

for i in range(CLUSTER_NUM):
    cluster_info['cluster' + str(i)] = df_kddcup[df_kddcup['cluster_id'] == i].mean()

cluster_info = cluster_info.drop('cluster_id')

plt.figure(figsize=(100,100))
kdd_plot = cluster_info.T.plot(kind='bar', stacked=True, title="Mean Value of Clusters")

kdd_plot.set_xticklabels(kdd_plot.xaxis.get_majorticklabels(), rotation=0)
plt.legend(bbox_to_anchor=(0., 1.02, 1., .102), loc=3,
       ncol=2, mode="expand", borderaxespad=0.)
```

Out[17]:
```
<matplotlib.legend.Legend at 0x1367d4e0>

<Figure size 7200x7200 with 0 Axes>
```